Development of Predictors for Cloud-to-Ground Lightning Activity using Atmospheric Stability Indices

THESIS

Kenneth C. Venzke, Captain, USAF

AFIT/GM/ENP/01M-8

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the U.S. Government.

AFIT/GM/ENP/01M-8

DEVELOPMENT OF PREDICTORS FOR CLOUD-TO-GROUND LIGHTNING ACTIVITY USING ATMOSPHERIC STABILITY INDICES

THESIS

Presented to the Faculty

Department of Engineering Physics

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the

Degree of Master of Science in Engineering and Environmental Management

Kenneth C. Venzke, B.S.

Captain, USAF

March 2001

Acknowledgments

I would like to express my appreciation to my faculty advisor, Lt Col Ronald P. Lowther, for his overwhelming insight, experience and unselfish willingness to provide expert guidance. Also, I would like to thank my committee members Maj Gary Huffines, Maj Thomas Reid, and Prof. Dan Reynolds, for their suggestions for improvement throughout the course of this thesis effort. Their insight and experience were certainly appreciated. I would also like to thank my biological mother's husband, of ARM Incorporated in Walnut Creek, California, whose insight, support and success through hard work has motivated me to strive to the "next level". The hard workers at the Air Force Combat Climatology Center, particularly SSgt John S. Kovachich and Mr. Peter W. Speck, should be praised for their outstanding support in the exhaustive database development provided to me to make this endeavor possible.

Finally, I express my appreciation to both my mothers (adoptive and biological), whose love and understanding over the past eighteen months have allowed me to dedicate many hours of research to complete this study.

Kenneth C. Venzke

Table of Contents

Page

List of Figures

Figure Page

List of Tables

AFIT/GM/ENP/01M-8

Abstract

A detailed examination was performed on several commonly applied atmospheric stability indices and lightning activity from 1993 to 2000 to determine the indices usefulness as predictive tools for determining cloud-to-ground lightning activity. Predetermined radii of 50 nautical miles around upper-air stations in the Midwest U.S. were used for the lightning summaries.

Also explored is an improvement upon the commonly accepted thresholds of the stability indices as general thunderstorm indicators. An improvement was found and new threshold ranges were developed for relating stability index values to lightning occurrence.

Traditional statistical regression methods failed to find a significant predictive relationship. By examining new techniques of data analysis, it was found that the detection and classification abilities of decision trees derived from the data-mining field best served the purposes of this study. Decision trees were examined on the large available database and significant results were found, resulting in the development of a lightning forecast tool for both the probability of lightning occurrence and its intensity. The predictive ability of the decision trees used in this study for lightning detection often exceeded 80-90% for most locations with a high degree of confidence.

The most significant features of the decision tree results were formulated into a forecast prediction tool with summary results for each location analyzed. These are specified both graphically and textually in a user-friendly format for forecasters to use as a "ready to use" predictive tool for forecasting lightning activity.

The results of this study using classification and regression trees were significant enough to implement immediately as a forecast tool for the operational weather forecast environment. Appendix A of this study is written as a "ready-to-use" forecast tool for weather forecasters. It is suggested that Air Force Weather units in the Midwest U.S. use this "innovative" forecast tool immediately for forecasting lightning activity.

DEVELOPMENT OF PREDICTORS FOR CLOUD-TO-GROUND LIGHTNING ACTIVITY USING ATMOSPHERIC STABILITY INDICES

I. Introduction

Thunderstorms with their associated lightning impact all aspects of military operations. For United States Air Force (USAF) weather forecasters, flight operations are most affected by lightning. For safety reasons, lightning activity in the area will halt most operations involving aircraft. Problems associated with lightning are not limited to Department of Defense (DoD) operations since many civil functions are significantly affected as well, such as agriculture, transportation, and especially the power/energy industry. The power industry relies heavily on thunderstorm forecasts, especially if significant lightning is anticipated. For example, inclement weather is the single largest cause of power outages, equating to as many as 40% of all interruptions. If thunderstorms are possible, great expenditure is made by this industry to put stand-by workers on call and get back-up generators started to minimize power interruptions. In addition, widespread cooling caused by evaporation during thunderstorm/rain events drastically reduces customer demand for air conditioning requirements during the summer season in the U.S. These effects are most significant in highly populated regions. Mismatches between generation capacity and customer demand either waste valuable resources or require expensive increases in supply for purchases of additional power at inflated wholesale prices (Dempsey et al., 1998).

Understanding and predicting thunderstorms and associated cloud-to-ground (CG) lightning activity in an operational environment proves both difficult and tasking, especially when considering the time constraints most operational forecasters operate under. This research examines atmospheric stability indices as possible predictive tools for CG lightning activity surrounding individual upper-air stations in the Midwest region of the United States.

1.1 Statement of the Problem

An upper-air station is a weather station that observes and disseminates weather balloon soundings from which the parameters for atmospheric stability indices are derived. Balloon soundings indicate the state of the atmosphere by measuring the temperature, humidity, and winds as functions of pressure (or height) as a balloon ascends up through the atmosphere. They are usually plotted manually or automated on a SKEW-T log-p diagram or in raw data format (AWS/TR-79, 1990).

Some of the most commonly calculated indices are the Lifted Index (LI), the K-Index (KI), and the convective available potential energy (CAPE). Seven indices were chosen and calculated for the locations used in this study. Operationally, it is usually left to the discretion of the forecaster to decide which index to use and which one is best representative for their region or particular weather regime.

Unfortunately, on particular days, certain indices may indicate severe potential, while on others they only indicate a slight risk of CG lightning activity. Both conditions tend to occur with varied results. This creates confusion as to the utility of the indices for the current forecast location or forecast region that is being examined. Experienced forecasters know that when analyzing the forecast environment for the potential of severe weather, the indices account for a large portion of the analysis and are a good starting point in the formulation of their forecast. However, this study shows that the indices specify a wide range of values for both days with CG lightning activity and days without any lightning activity at all.

For very active CG lightning days, which may or may not be associated with severe weather at the surface, there is thought to be a noticeable relationship to a limited range of unstable index values. The significance of this relationship has been the focus of studies accomplished on stability indices in the past, but never with substantial justification (Coleman, 1990). This study attempts to definitively assess some of the most common indices used in operational weather forecasting and, to ultimately develop forecast tools in which these indices are suitable as predictors of CG lighting activity for individual locations or regions in the Midwest.

Experienced forecasters seem to have their "favorite" stability index, but unfortunately forecasters are unable to determine which stability index to rely upon the most for every weather regime being forecasted for. Furthermore, even experienced forecasters should not rely totally on just one of the stability indices for their forecast and may not want to even consider using them at all under certain conditions.

Stability indices have historically been used to assess the threat and potential severity of thunderstorms (with which CG lightning activity is clearly associated). However, it appears no previous studies have assessed the degree to which stability indices may be used as predictive tools for CG lightning activity or its intensity as provided by the highly dependable and proven accuracy of the National Lightning Detection Network (NLDN), especially over the Midwest region of the United States.

The goal of this research then, is to ascertain the best relationship possible between stability indices for use as forecast tools in predicting any CG lightning or the amount of activity surrounding upper-air stations in the Midwest. Any predictive relationships found will increase weather forecaster's confidence levels in their use and ability to predict CG lightning activity. Any increase in the ability to accurately predict CG lightning events and activity will be beneficial to all DoD and civil operations affected by CG lightning activity.

1.2 Research Objectives

In the absence of adequate predictive tools for forecasting CG lightning events, this study examines the use of atmospheric stability indices as a means to discover methods to exploit any possible significant relationships. The specific tasks necessary to achieve the goal of this study were:

1. to determine the most useful radii (10nm, 25nm, or 50nm) of CG lightning summaries around a station to examine relationships with and to combine the most homogeneous months of lightning activity to maximize the usefulness of the dataset for each upper-air location;
2. to analyze the stability indices and formulate an improved range of values to combine with CG lightning occurrences;
3. to examine the CG lightning data and stability indices for any predictive relationships by using statistical regression (linear and non-linear) techniques;
4. to exploit data mining techniques to introduce new predictive techniques and to establish the most significant threshold values among the stability indices using the detection and classification abilities of decision trees; and,
5. to formulate a forecast matrix using any predictive relationships found of the most

significant features as determined by the decision tree results.

II. Background and Literature Review

2.1 *Lightning Background*

The lightning activity data used in this study indicates cloud-to-ground (CG) strikes only within a 50 nautical mile (nm) radius of each predefined upper-air station in the Midwest, as disseminated by the National Lightning Detection Network (NLDN). Relationships were made between CG strikes at different radii (50nm, 25nm, and 10nm) in twelve-hour (12Z to 00Z and 00Z to 12Z) increments to coincide with matching upper-air sounding times for representation. It was quickly determined that the lightning data at 50nm was most representative for lightning in the general vicinity of a station. The CG lightning strike summaries for the 25nm and 10nm radii seemed to capture too few occurrences and therefore less significant relationships could be inferred between the indices and CG strike activity.

A radius of 50nm was chosen to represent the atmosphere around each upper-air station for comparison reasons and was used as the starting point to assess any potential utility of the stability indices for predicting CG lightning activity. CG lightning is continuously referred to because, as will be shown later, lightning activity data from the NLDN consists of CG lightning strikes only. No intra-cloud lightning measurements are inferred due to limitations in sensor threshold measurements (Cummins et al., 1998).

There are limitations to the lightning data used in this study. Progress in detecting CG lightning strikes has been well documented in a recent publication by Cummins et al. (1998), who summarized the detection efficiency of the network from its past to its present form. Prior to 1992, GeoMet Data Services (GDS), the organization that maintained the network during that time, estimated that the average location accuracy of CG lightning strike locations varied from 8 to 16 km in the NLDN. The flash detection efficiency during this same period was around 70%, using first stroke peak currents of greater than 5 kiloAmps (kA). Data model estimates do not include flashes with peak currents less than 5kA and are not considered a CG flash because of large uncertainties in the peak current distribution at lower amperages. In early 1992, GDS calibrated the sensors, increasing the accuracy of the network to 4 to 8 km, with a flash detection efficiency of 65 to 80%. Once an upgrade in 1995 was completed, the location accuracy improved to 1 to 2 km, with a first stroke detection efficiency of 80 to 90%. However, manual video verifications showed detection efficiencies of 84% prior to the upgrade in 1994 and 85% detection efficiencies in 1995 after the upgrade. Therefore, significant ambiguities between data prior to 1995 are not expected and appear acceptable (Wacker and Orville, 1999).

Prior to the establishment of the NLDN, documenting thunderstorm events was through visual observations or perhaps radar and satellite information to supplement detection. The most significant deficiency of this system is the timeliness and accuracy of reporting. The NLDN alleviates these potential inaccuracies by providing automated near real-time reporting of CG lightning data to forecasters. Since 1991, upgrades to NLDN sensors have increased the accuracy of stroke detection significantly. The most recent upgrade, completed in 1995, reduced the total number of sensors from 130 to 106 because of an increase in the effective range of the existing sensors (Cummins et al., 1998). The location accuracy has been improved by a factor of 4 to 8 since 1991, resulting in a median location accuracy of approximately 500 meters at its best. The detection efficiency increased from 65-80% in 1992-1994 to 80-90% after the 1995 upgrade. This is significant since most stability indices and lightning data used in this study included the 1993-1994 period of record.

However, there were a few locations where only data since 1995 were available (see Table 1).

Table 1. Data availability for each location used in this study.

WMO	ICAO	Location	State	Elev (m)	Lat		Lon		Period of Record
72248	SHV	SHREVEPORT REGIONAL	LA	79	32.28	N	93.49	W	2/95-5/00
72249	FWD	FORT WORTH	TX	196	32.50	N	97.18	W	7/94-5/00
72340	LZK	NORTH LITTILE ROCK	AR	165	34.50	N	92.15	W	1/93-5/00
*72355	FSI	FORT SILL (Military)	OK	362	34.39	N	98.24	W	1/93-5/00
72357	OUN	NORMAN/WESTHEIMER	OK	357	35.13	N	97.27	W	1/93-5/00
**72363	AMA	AMARILLO ARPT(AWOS)	TX	1099	35.14	N	101.42	W	1/93-5/00
72440	SGF	SPRINGFLD MUNI(AWS)	MO	387	37.14	N	93.23	W	5/95-5/00
72451	DDC	DODGE CITY(AWOS)	KS	790	37.46	N	99.58	W	1/93-5/00
72456	TOP	TOPEKA/BILLARD MUNI	KS	270	39.04	N	95.37	W	5/95-5/00
72558	OAX	OMAHA/VALLEY	NE	350	41.19	N	96.22	W	7/94-5/00
72562	LBF	N. PLATTIE/LEE BIRD	NE	849	41.08	N	100.41	W	1/93-5/00
72662	RAP	RAPID CTY RGNL ARPT	SD	964	44.05	N	103.03	W	1/93-5/00
74455	DVN	DAVENPORT UPPER-AIR	IA	229	41.37	N	90.35	W	3/95-5/00

* 12Z sounding only

** SWEAT index missing 11/98-5/00

2.2 Stability Index Background

Weather balloons attached to their Styrofoam-boxed instrumentation called rawindsondes have been used to gather atmospheric measurements of the vertical temperature, moisture, and wind profiles (soundings) above a location since the early 1900s. Rawindsondes have been the foundation of the global upper-air observing system with more than 1,000 rawindsonde stations operated by 92 countries as of the early 1990s (NOAA, 1992). Most of these upper-air stations in the United States launch weather balloons twice a day, once at 00Z (Universal Time Coordinated (UTC) or Greenwich Meridian Time (GMT) and again at 12Z. Across the continental United States, weather balloons are launched from over 100 different locations, from which many various calculations are made from the environmental data gathered. These range from the complex analysis/forecast models developed by weather organizations to the derived stability indices used in this research effort. Due to the inaccuracy, at times, of these weather models, research to improve them is a continuous effort. Stability indices, then, are an essential part of the analysis/forecast process (especially for convective weather forecasting) and are used in combination with the analysis/forecast models to determine the current and forecasted states of the ever-changing atmosphere. Thirteen sounding locations in the Midwest were chosen for this study from the various government and military sounding sites indicated in Figure 1. While the results of all 13 locations are presented, this study focuses on two of the sites, which are deemed representative of the entire regional climate regime. One in Oklahoma, a National

Weather Service (NWS) sounding site, is Norman (OUN) and the other, in west-central Nebraska, is North Platte (LBF).

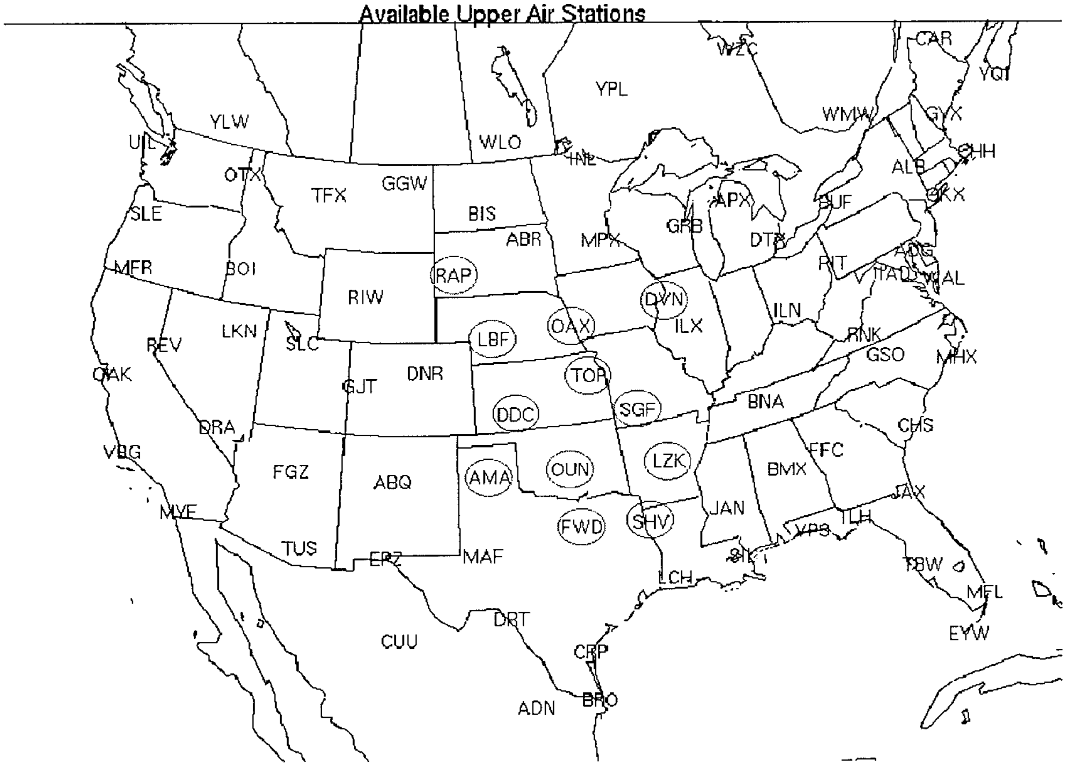


Figure 1. United States upper air stations along with their corresponding ICAO (International Civil Aviation Organization) identifiers. The Midwest sounding sites included in this study are circled.

2.3 *Atmospheric Sounding Data Reliability*

The soundings, derived from the rawindsondes discussed previously, refer to a profile of vertical distribution (from a single location) of the pressure, temperature, dew point temperature, wind direction, and wind speed from measurements taken by a rawindsonde as it traverses upward near the site where the balloon was launched. Depending on the strength of the winds aloft though, the information gathered is usually not representative of the atmosphere immediately over the launch site. It would be ideal to have an exact replication of the current state of the atmosphere directly above each measurement location. However, because strong winds aloft blow the balloon a considerable distance downstream from where it was released, this is usually not the case. The measurements though must be considered representative of the sounding location, because no location error corrections are made to rawindsonde observations (Andra, 2000). The location errors are especially large when there are strong upper-level winds blowing the sounding balloon further away from the launch site as it rises into the atmosphere. This makes the data even less representative of the location from which it originated. Fortunately, most of the indices calculated for this study compute temperature and moisture measurements from the 850 and 500 millibar (mb) pressure levels, which equate to roughly 3,000 to 18,000 feet, respectively, in the standard atmosphere. Rawindsondes typically take measurements well above 300mb (over 30,000 feet), where location errors can be quite large. For purposes of this study, it is assumed, as it is for the national rawindsonde network, that these errors are minimal and thus are not considered significant, especially in a data dense region such as the Midwest with diminutive terrain complexity.

2.4 Stability indices as Predictors

There are numerous other severe weather indices in use, many of which are used at the National Severe Storms Forecast Center by forecasters who specialize in severe weather forecasting. The indices presented in this research effort are those routinely used by forecasters to evaluate the stability or instability of the atmosphere. The stability indices can be thought of as the analyzed convective potential of a sounding expressed as a single numerical value. Miller et al. (1972) developed the generally accepted stability index thresholds for the Midwest that were used in this study. The stability indices were further classified into the threshold categories listed in Table 2. From these categories, an improved range of values for the occurrence of CG lightning is suggested in the next chapter. But first calculations of each stability index are discussed.

Table 2. Suggested range of index values as general thunderstorm indicators (AFWA, 1998).

Index	REGION best applied	Weak (Low)	Moderate	Strong (High risk)
CAPE	East of Rockies	300 to 1000	1000 to 2500	2500 to 5300
K-Index	East of Rockies in moist air	20 to 26	26 to 35	> 35
KO-Index	Cool, moist climates (Pacific	> 6	2 to 6	< 2
Lifted Index	All	0 to 2	-3 to -5	< -5
Showalter	CONUS	> 3	2 to -2	< -3
Total Totals	East of Rockies	44 to 45	46 to 48	> 48
SWEAT (for Severe)	Midwest and Plains	< 275	275 to 300	> 300

2.5 Stability Index Calculations

Convective Available Potential Energy (CAPE)-

CAPE is a measure of the amount of buoyant energy available to accelerate a parcel of air vertically. CAPE is directly related to the maximum potential vertical speed within an updraft or a summation of the amount of buoyancy (not accounting for drag or non-adiabatic effects). Higher values indicate greater potential for severe weather. Observed values in thunderstorm environments often exceed 1,000 joules per kilogram (J/kg), and in extreme cases may exceed 5,000 J/kg. However, as with the other indices, a wide range of values are associated with a wide range of weather phenomena, notwithstanding, lightning activity. CAPE is represented on a skew-T log-P diagram as the area of energy enclosed between the environmental (sounding) lapse rate and the parcel derived lapse rate from the LFC (Level of Free Convection) to the EL (Equilibrium Level) (AWS TR-79/006, 1979). This area, often called the positive area, is directly related to positive buoyancy. This positive area represents the maximum potential strength of updrafts within a thunderstorm, should one develop.

CAPE values of greater than 1,500 J/kg, dependent upon location and season, represent enough energy to produce thunderstorms. A value greater than 3,000 J/kg represents enough energy to produce strong thunderstorms. Negative values of CAPE denote a relatively stable atmosphere and are referred to as Convective Inhibition (CIN), which is computed as the negative area on the sounding, if it exists (AWS TR-79/006, 1979). CIN was not computed for this study. Knowledge of a CAPE profile or the shape of a sounding also has some implications, but was not considered. For example, two soundings may have the same CAPE values but different profile shapes (South African Weather Bureau, 2000). This study therefore utilizes the positive values of CAPE in comparisons with CG lightning activity.

Showalter Stability Index (SSI)-

The SSI (Showalter, 1953) is a measure of the potential instability in the 850mb to 500mb layer. The SSI may be unrepresentative if significant amounts of moisture reside below 850mb with dry air residing above. In this case the SSI would not be able to detect the resulting instability. SSI is the stability index most commonly used by military and other forecasters. It indicates the general stability of an air mass but should not be used when a frontal boundary or a strong inversion is present between the 850mb and 500mb levels. SSI is computed using the layer between 850mb and 500mb as follows:

$$SSI = T500 - TP500 \quad (1)$$

where,

- T500 = the measured temperature in degrees Celsius at 500mb
- TP500 = temperature in degrees Celsius of an air parcel lifted moist adiabatically from the 850mb lifted condensation level to 500mb

Lifted Index (LI) -

The LI (Galway, 1956) is a measure of the potential instability from the surface to the 500mb level. It is very similar to the SSI, but instead of using the arbitrary choice of the 850mb level, it is usually computed by lifting a parcel with an average mixing ratio along the dry adiabat in the lowest 3,000 feet of the sounding using the mean mixing ratio by equal area averaging to better consider the available low-level moisture below the 850mb level. There are various methods used to determine the initial level. Some methods use the maximum forecasted afternoon temperature or the mean sounding temperature in the lower levels if significant heating or cooling is not expected during the afternoon. The algorithm used at the Air Force Combat Climatology Center (AFCCC) uses the average mixing ratio in the lower 3,000 feet to compute the LI for this study.

A common measure of atmospheric instability, the LI is obtained by computing the temperature that air near the ground would have if it were lifted to 500mb (approximately 18,000 feet for the standard atmosphere) and comparing that temperature to the actual temperature at that level. Positive values reflect stable conditions while negative values reflect unstable conditions (the parcel is warmer than its environment so it will continue to rise. It is computed as follows:

$$LI = T(500\text{mb environment}) - T(500\text{mb parcel}) \quad (2)$$

The LI is measured in degrees C, where "T(500mb environment)" represents the 500mb environmental temperature and "T(500mb parcel)" is the rising air parcel's 500mb temperature. If the lifted air parcel is warmer than its surrounding environmental temperature then it should continue to rise. Thus, negative values indicate instability and the more negative, the more unstable the air is, and the stronger the updrafts are likely to be with any developing thunderstorm(s).

Total Totals Index (TTI)-

The TTI (Miller, 1972) consists of two components: Vertical Totals (VT) and Cross Totals (CT). VT represents static stability between the 850mb and 500mb levels while the CT includes a moisture parameter, the 850mb dew point temperature. As a result, TTI accounts for both static stability and 850mb moisture amounts. However, TTI can be illusory in situations where the low-level moisture may reside below the 850mb level. For example, if a significant capping inversion is present, convection may be inhibited even when TTI values are strong.

TTI, like SWEAT (described next), is actually a compound index designed to better predict the occurrence of severe weather, not just general thunderstorms. In other words, it was developed for use when such indices as SSI or LI indicate that thunderstorms may occur. However, this index is another more commonly derived index that many novice weather experts may assess equally along with SSI and LI to determine relative instability. This is why all the commonly derived indices were used in this study. Additionally, it is unknown whether any of these indices have a predictive relationship to CG lightning strikes within 50nm of a station. It appears that the predictive potential of the indices to CG lightning activity within a specified radius of a sounding location has never been assessed. It will be seen later that in fact some of the indices developed specifically for severe weather indication appear to correlate well to CG lightning counts. TTI is computed as follows:

$$TTI = (T850 - T500) + (D850 - T500) \quad (3)$$

To calculate the TTI, two values are computed from the sounding: the vertical totals (VT) and the cross totals (CT). VT is a measure of the vertical stability without regard for moisture parameters and is computed by subtracting the 500mb temperature (T500) from the 850mb temperature (T850). CT is a measure of stability that includes moisture and is found by subtracting T500 from the 850mb dew point temperature (D850). The Total Totals (TTI) index is simply the sum of VT and CT. Forecasters evaluate thunderstorm potential according to the general guidelines provided by Miller (1972). The TTI index is the most reliable single predictor of severe activity for both warm and cold seasons. During 1964 and 1965, 92 percent of all reported tornadoes occurred with a TTI of 50 or greater. Most widespread tornado outbreaks occurred with a TTI of 55 or greater (Miller et al., 1972). High values of TTI can result with insufficient low-level moisture (determined by CT), which is required for convective activity. In other words, a low CT combined with extremely high VT values can suggest misleading TTI values. This is another reason to integrate other indices into a forecasters "convective potential equation". Other indices account for various other temperature and moisture parameters that the TTI may miss with its single consideration for moisture at the 850mb level.

TTI must be used with careful attention to either the CT value or the actual low-level moisture amounts, since it is possible to have a large TTI value with insufficient low-level moisture to support thunderstorms.

Severe Weather Threat Index (SWEAT)-

The SWEAT Index (Miller et al., 1972) evaluates the potential for severe weather by examining both kinematics (wind) and thermodynamic information into one index. It is one of the more complex indices derived in this study, resulting in this index as having one of the highest missing data rates. The algorithm used by AFCCC to compute this index requires wind parameter measurements at specified height levels. If any of these required measurements are missing, then the index cannot be calculated. These parameters include low-level moisture (850mb dew point), instability (via TTI), lower and middle-level (850 and 500mb) wind speeds, and warm air advection (veering between 850 and 500mb). Unlike KI, the SWEAT index was originally developed to assess severe weather potential, not just ordinary thunderstorm potential.

$$\text{SWEAT}=(12*850\text{Td})+(20*[\text{TTI}-49])+(2*\text{f}850)+\text{f}500+(125*[\text{s}+0.2]) \quad (4)$$

where,

- 850Td is the dew point temperature at 850mb,
- TTI is the total-totals index,

- f850 is the 850-mb wind speed (in knots),
- f500 is the 500-mb wind speed (in knots), and
- s is the sine of the angle between the wind directions at the 500mb and 850mb levels (thus representing the directional shear in this layer) which equates to the amount of warm air advection between the layers.

The last term in the equation (the shear term) is set to zero if any of the following criteria are not met:

1) 850mb wind direction ranges from 130 to 250 degrees,

2) 500mb wind direction ranges from 210 to 310 degrees,

3) 500mb wind direction minus the 850mb wind direction is a positive number, and

4) both the 850 and 500mb wind speeds are at least 15 knots.

No term in the equation may be negative; if so, that term is set to zero.

Guidance values developed by the Air Weather Service suggest severe storms may be possible for SWEAT values of 250-300 if strong lifting is present. In addition, tornadoes may occur with SWEAT values below the 400mb level, especially if convective cell and boundary interactions increase the local shear, which cannot be resolved in this index. The SWEAT value can increase significantly during the day, so low values based on 12Z soundings may be unrepresentative if substantial changes in moisture, stability, and/or wind shear occur during the day. SWEAT values of about 250-300 indicate a greater potential for significant thunderstorms, but as with many of the stability indices, there are no significant "magical" thresholds developed for CG lightning activity.

K-Index (KI)-

The K index (George, 1960) or K Value is a measure of thunderstorm potential based on the vertical temperature lapse rate along with the amount and vertical extent of low-level moisture in the atmosphere. The KI is computed as follows:

$$KI = T850 + D850 - T700 + D700 - T500 \quad (5)$$

KI is a measure of thunderstorm potential based on the temperature lapse rate, the moisture content of the lower atmosphere, and the vertical extent of the moist layer. It should be used to analyze the potential for air mass thunderstorm occurrence—not potential occurrences of frontal thunderstorms and not for the potential severity of a thunderstorm. The temperature difference between the 850mb and 500mb heights is the parameter used to find the vertical lapse rate, and the 850mb dew point and the 700mb dew point depression are used to evaluate the moisture content of the air, as well as the vertical extent of the moist layer.

As was mentioned earlier, each index has its own advantages and disadvantages. The main weaknesses depend primarily upon the levels of analyses used for each index or the shape of the atmospheric profile.

III. Data Collection and Review

It is important to appreciate the history, background, and potential for weaknesses of the data used in this study. Basically, there are two separate sources of data, lightning summary output derived from the NLDN, and stability indices derived from the moisture, wind, and temperature profiles of upper air soundings. Each stability index is a measure of the potential instability of the atmosphere by an examination of the different combinations of temperature and moisture at pre-determined pressure levels (or heights).

A more formal definition of a stability index is: the analyzed convective potential of an upper-air sounding expressed as a single numerical value. The importance of the stability indices was pointed out by a study conducted by Air Weather Service and the National Severe Storms Forecast Center of the National Weather Service (Miller, 1972). This survey used 328 tornado cases to determine which atmospheric conditions were necessary for the development of severe weather. The parameters were ranked in order of importance based on both computer analysis and forecasting experience. An analogy is drawn to this approach later in this study using data mining techniques. Results showed that the second most influential parameter for convective forecasting is the stability of the atmosphere itself, upon which the applications of the stability indices are based. Additional stability indices have been developed since the original study (Miller, 1972) and are considered in this study as well. Each index takes different atmospheric parameters into consideration. Not considered, but readily available from upper-air soundings, are the wind field structures and the indices derived from them, such as helicity or upper-level flow.

3.1 Data Methods

In an attempt to improve weather forecasts for cloud-to-ground (CG) lightning activity, which is inherently related to thunderstorm convection, stability indices and CG lightning relationships were examined for 13 different upper air stations in the Midwest. Again, no inferences are made at this point between severe or non-severe types of convection. An exhaustive effort was made to utilize a large sample database of upper air soundings in which the indices are derived for each location along with highly accurate CG lightning summaries from the NLDN between 1993 through 2000. Relationships were made between CG strikes at different radii (50nm, 25nm, and 10nm) in twelve-hour increments (12Z to 00Z and 00Z to 12Z) increments to coincide with matching sounding times for representation. It was quickly determined that the lightning data for 50nm model was the most representative for an area around each station. The CG lightning strike summaries for the 25nm and 10nm radii seemed to capture too few occurrences and therefore no relationships were inferred between these indices and CG strike activity.

The average horizontal spacing of the upper-air sounding network in the Midwest is approximately 200nm. However, the summer environment in the Midwest is often characterized by shower-producing systems that occur on smaller spatial scales. These may be missed by the 50nm radius used in this study for CG lightning strike comparisons, yet may still be representative of the general area around the sounding. This fact could potentially degrade the significance of any results since many CG lightning strikes may be missed.

When one looks at the time scales of convective activity in this region during the summer months, many times they are on the order of only a few hours. Thus, in a rapidly changing environment, such as storms triggered by a fast moving frontal system, these sounding analyses may miss key moisture and temperature changes affecting the stability of the atmosphere.

Some potential limitations to this approach, based on past research, can be anticipated. The goal in this study is to predicate past studies that utilized the stability indices but had proven inconclusive (Huntrieser et al., 1996). Alternatively, a study was conducted in the High Plains with a high-resolution mesonet with 25-50km spacing and twice the number of sounding observations than are normally available on a day-to-day basis (Mueller et al., 1993).

The results indicated that, much to their dismay, a high resolution of timely mesonet upper-air soundings provided no further skill of the soundings to predict convective weather outbreaks. Therefore, their conclusion was that the existing sounding observation resolution should be adequate for research purposes. This point is very important, since it helps further justify the available sounding databases used in this study.

Parameters such as helicity, streamwise vorticity, and hodographs have proven results when combined with atmospheric stability indices but require a bit more detailed analysis then was able to be considered in this study (AWS TR 79-006, 1990).

3.2 Data Sources

The Air Force Combat Climatology Center (AFCCC) located in Asheville, NC provided the upper-air stability indices and the NLDN summary data used in this study. AFCCC has an extensive database of NLDN lightning data and raw upper air data for the Midwest with the ability to provide the data in many different formats.

Lightning summary data for CG strikes within a 50nm radius of each location for each 12 hour period were calculated using archived NLDN data and ArcView GIS mapping applications. Once the raw data was formatted, ArcView easily determined the daily counts. Typically it took one day per location to complete the summaries.

The stability indices utilized in this study were computed using archived upper-air sounding data ingested by FORTRAN algorithms developed at AFCCC. These were much easier to compute than the lightning summaries. Unfortunately, algorithms were not available for every stability index and time constraints prohibited the development of new algorithms to create any additional indices needed. Suggested applications of other indices not considered in this study are recommended for future consideration in the last chapter of this study.

IV. Methods of Data Analysis

4.1 Analysis of stability indices in deciphering homogeneous datasets

Box plots and histograms of the indices and CG lightning data were constructed for each location with results displayed for LBF (North Platte, NE) and OUN (Norman, OK). A more thorough attempt was made to analyze the datasets for display at LBF and OUN because of their large available datasets (1993-2000) and representative locations in the northern and southern portions of the study (Figure 2).

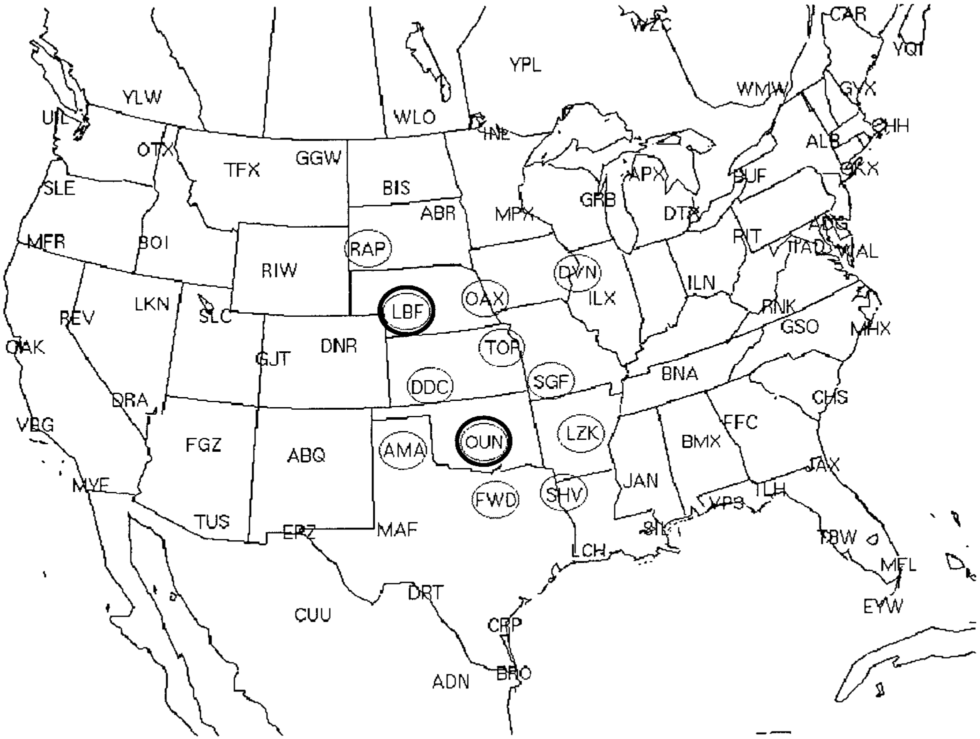


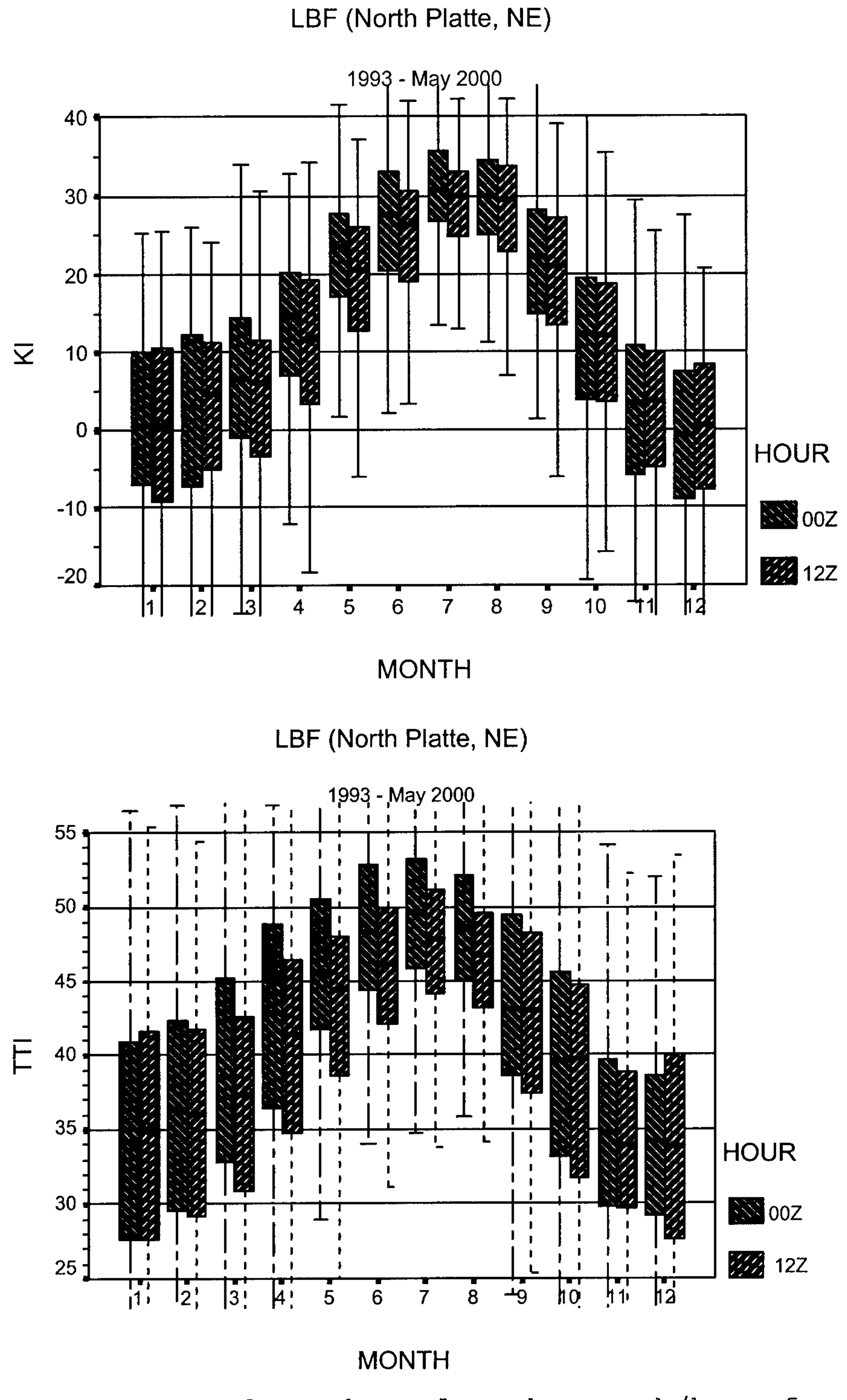
Figure 2. Locations used for this study with emphasis on LBF and OUN.

Determining which months are homogeneous in this study includes a month-by-month assessment of the available data. It was ascertained that certain potentially unreliable non-homogeneous datasets should be eliminated and the rest combined. Combining the significant months helped maximize the database for each location.

In Figures 3-6, not surprisingly, a noticeable peak in the summer months for all locations toward more unstable values of selected indices is evident for both sounding times (00Z and 12Z). The summer months from May to September (5-9) project the peak instabilities the most. Note that only positive values for CAPE are used.

There is also a noticeable increase in the variability (range of values) between 00Z and 12Z of the indices shown in Figures 3-6. It is clearly evident the effects that morning inversions have on 12Z sounding times during the "active" season. The 12Z CAPE calculations are especially variable because of the way it is calculated (integration through the atmosphere which can be concealed at inversion levels). It is determined shortly that a more unstable range of values for most indices is required for the 12Z soundings to be associated with CG lightning activity. It is not surprising that the afternoon 00Z soundings appear to be most representative when convection is possible or expected.

Mueller et al. (1993) determined that forecasted afternoon soundings correlated best to thunderstorm activity. In their study, the 12Z forecast soundings performed better than the 00Z soundings. Forecast soundings were considered in this study but forecast upper-air model data are not archived in any known data center and time limitations prohibited their development. Under that rationale, forecast soundings for each location are suggested for future research.


LBF (North Platte, NE)

1993 - May 2000

LBF (North Platte, NE)

1993 - May 2000

Figure 3. KI and TTI box plots by month/hour for LBF, showing seasonal trends and peak instability for months 5-9.

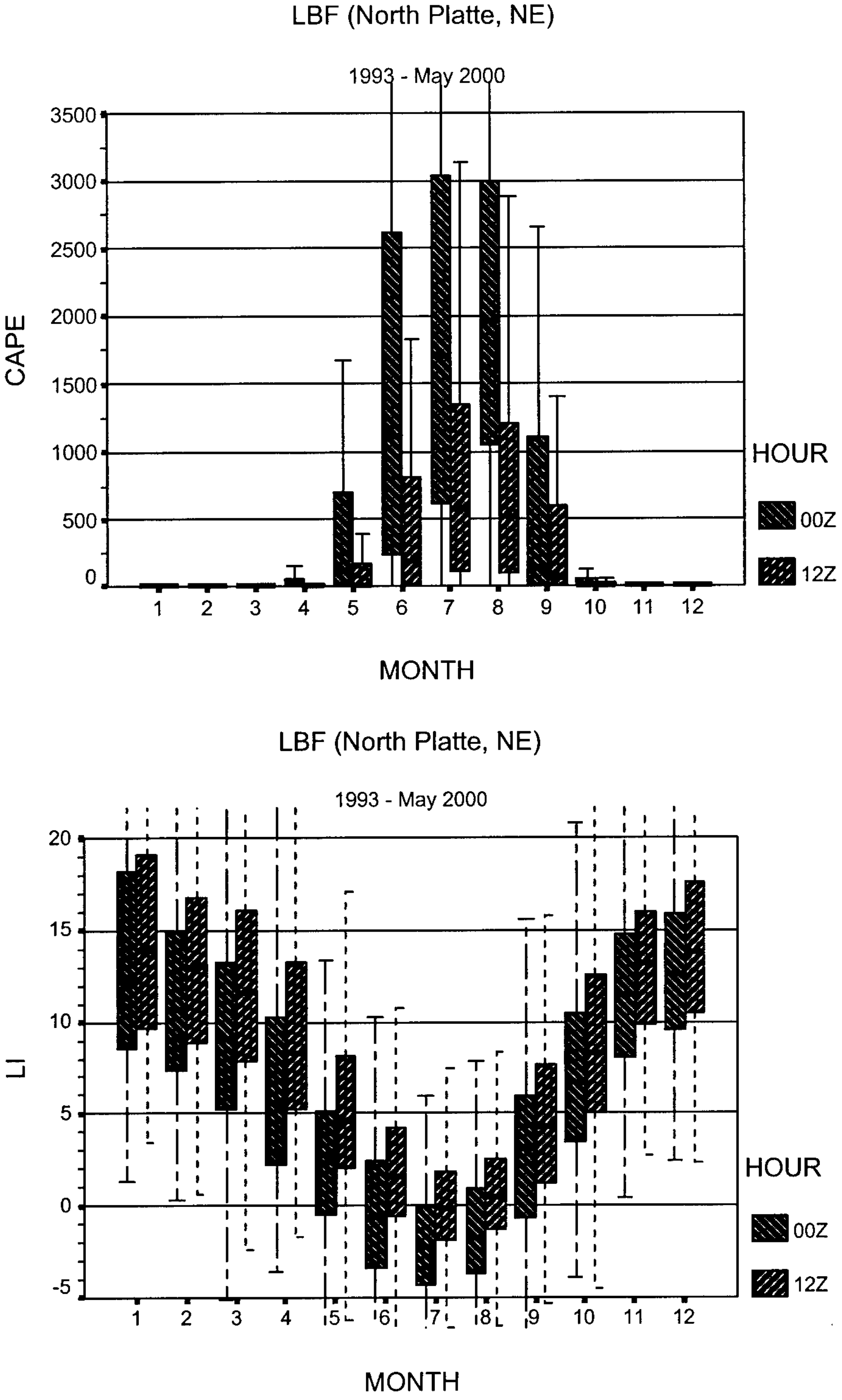


Figure 4. CAPE and LI box plots by month/hour for LBF, showing seasonal trends and peak instability for months 5-9.

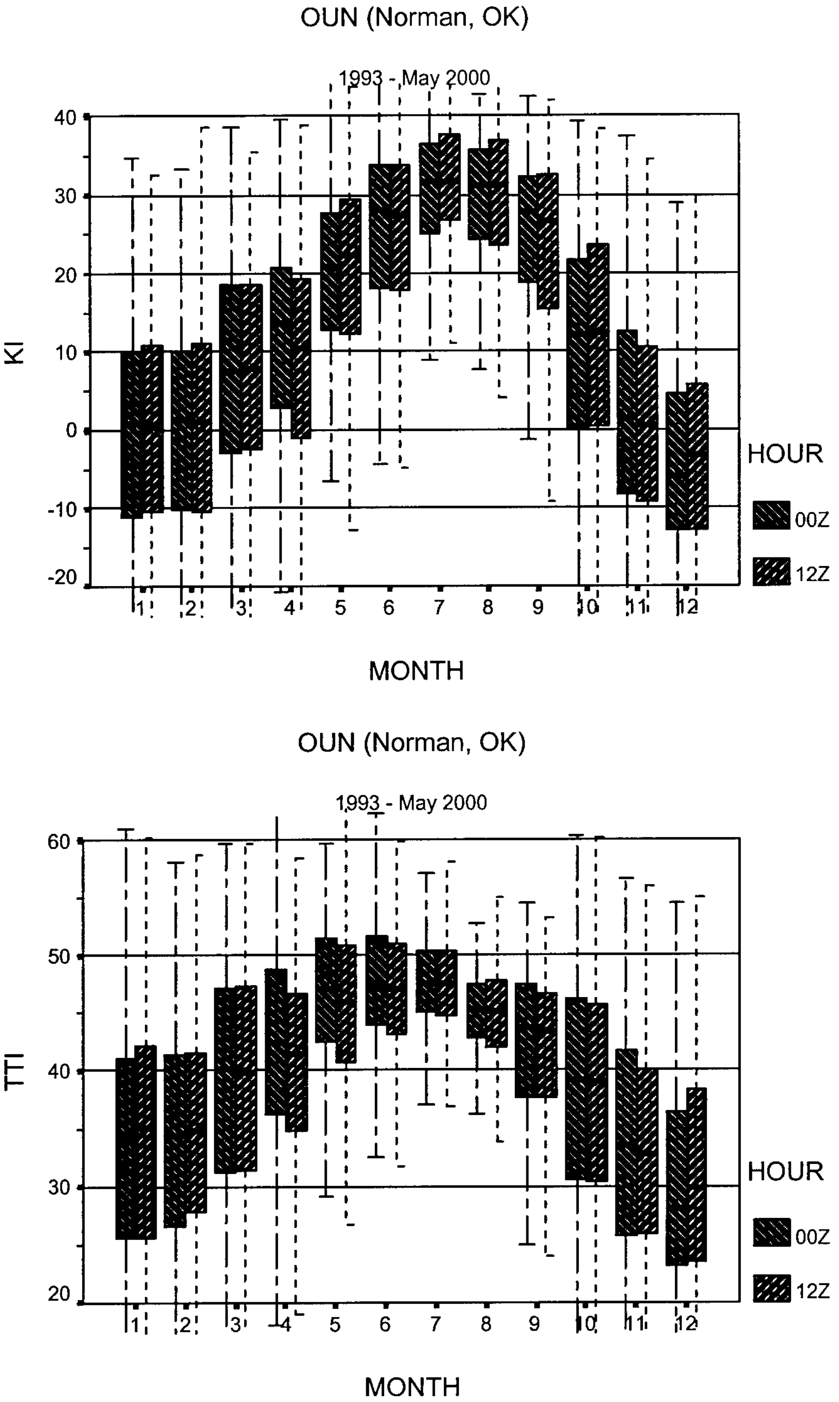


Figure 5. KI and TTI box plots by month/hour for OUN, showing seasonal trends and peak instability for months 5-9.

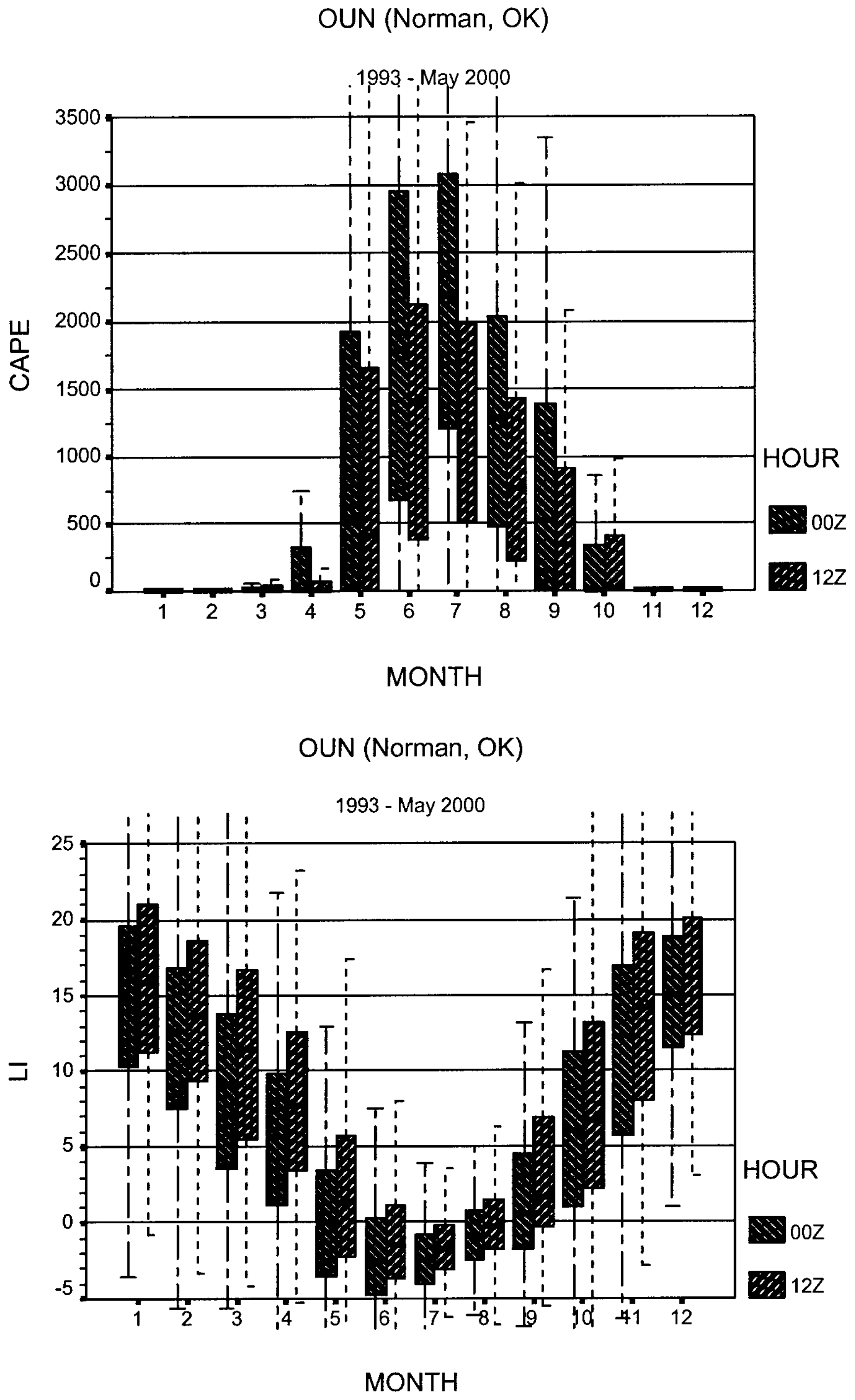


Figure 6. CAPE and LI box plots by month/hour for OUN, showing seasonal trends and peak instability for months 5-9.

4.2 Analysis of Cloud-to-Ground Lightning Activity in Deciphering Homogeneous Datasets.

Bar plots of mean monthly CG lightning activity also show a distinct peak during months 5-9 for 00Z and 12Z at both locations (Figures 7 and 8). Bar plots for the total number of days with any CG lightning activity (labeled CG_COUNTS) were also constructed in Figures 9 and 10. It may be arguable whether to include months 4 or 10 at 12Z for OUN as the active lightning months, but LBF definitely supports the hypothesis that months 5-9 are the most active for CG lightning activity and would exhibit the most homogeneity for all locations.

4.3 Maximizing the Datasets

To optimize the usefulness of the database for each location, an effort was made to see if it would be reasonable to merge specific "active" months together for analysis purposes. The indices and CG lightning summaries obviously show a significant variability by season. CG lightning counts are significantly lower or non-existent and index values indicate much higher variability for the "cool" or inactive months from October to April. Conversely, significant peaks in CG lightning counts and less variability for most indices existed for months 5-9.

Accordingly, to further optimize the usefulness of the dataset for each location, it was decided to combine the more homogeneous data set of just the warm "most active" months (5-9) for determining underlying threshold values of the indices to CG lightning activity.

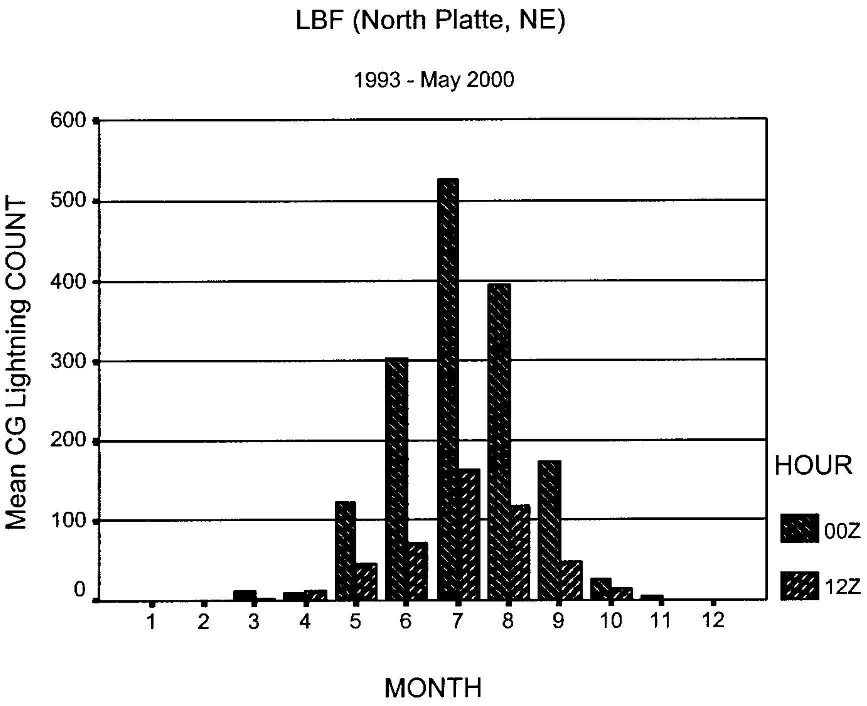


Figure 7. Mean CG count by time/month for LBF.

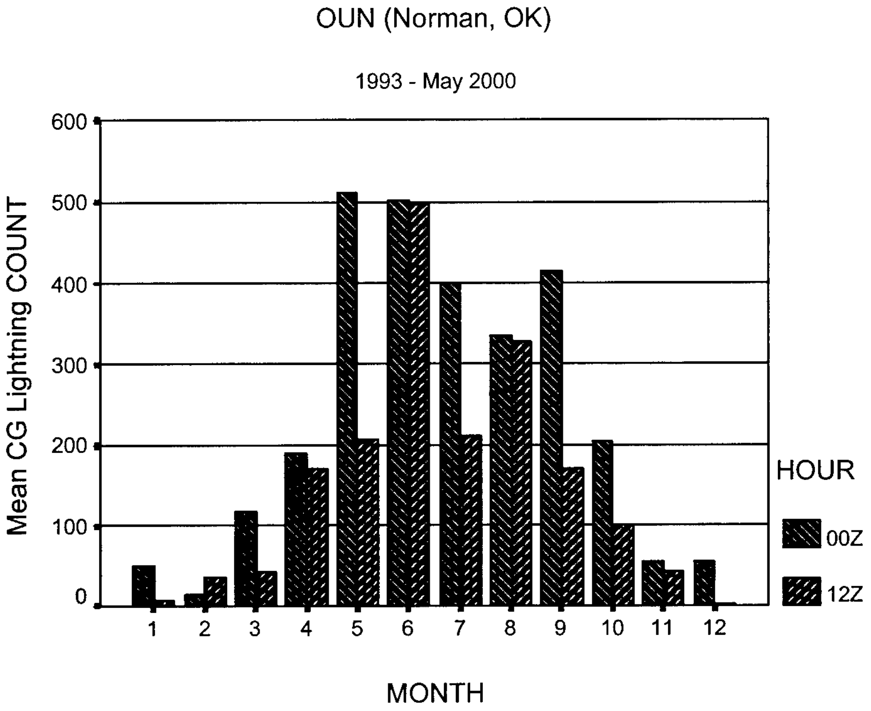


Figure 8. Mean CG count by time/month for OUN.

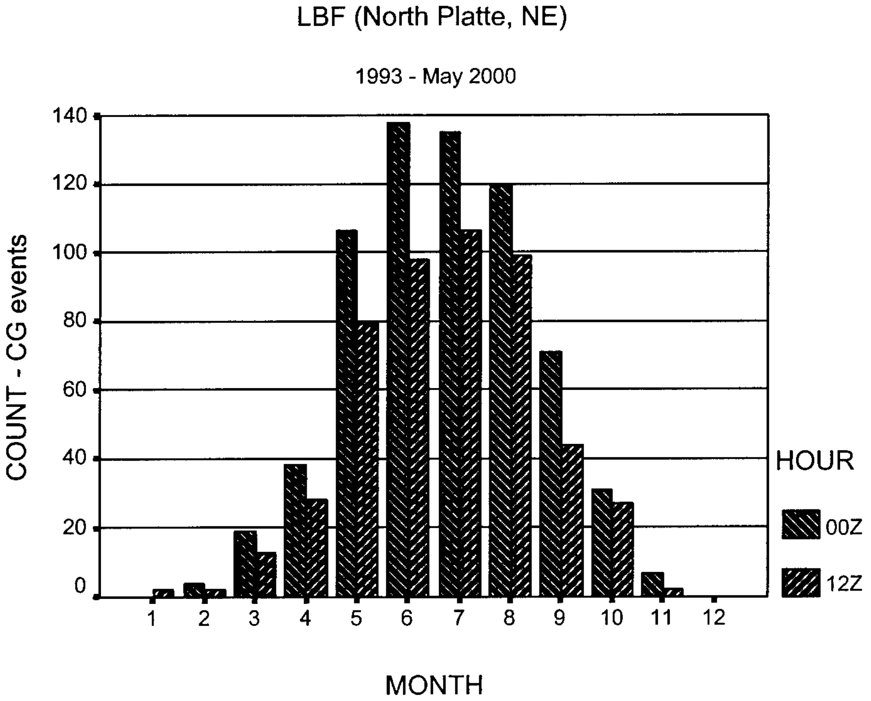


Figure 9. Total CG lightning days within a 50nm radius of LBF.

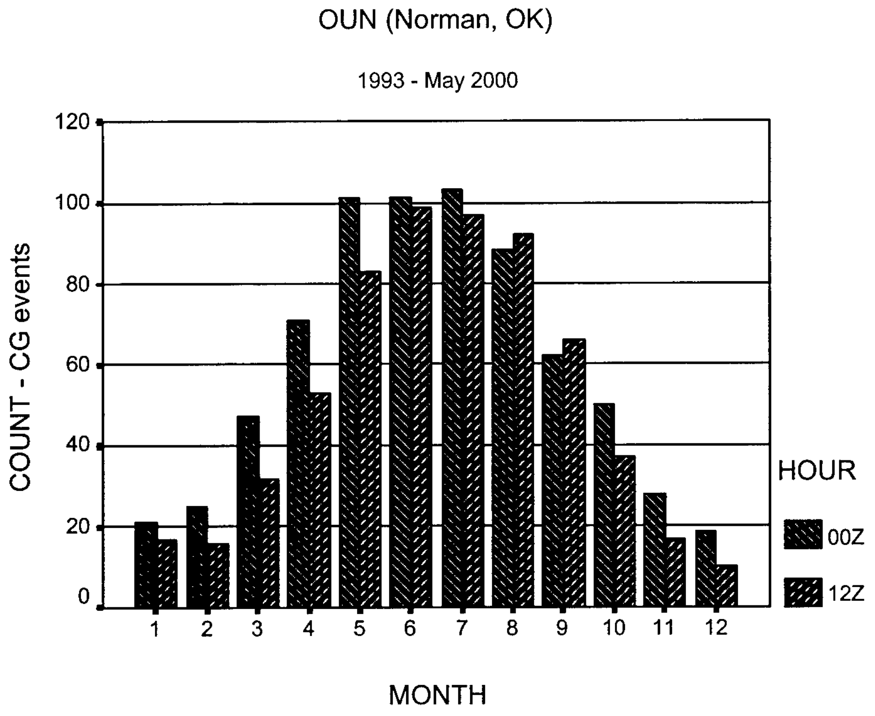


Figure 10. Total CG lightning days within a 50nm radius of OUN.

4.4 Developing a Baseline Climatology of Stability Index Values for Predicting CG Lightning Activity.

In their article "A baseline climatology of sounding-derived supercell and tornado forecast parameters", E. Rassmussen and D. Blanchard discuss the need for a baseline climatology of sounding threshold values to weather events in support of operational thunderstorm forecasts. Their study concentrated primarily on the climatology of CAPE and other more dynamic weather parameters to severe thunderstorm and tornado occurrences. The question that they felt needed to be answered was "at what values or thresholds of stability indices should forecasters become concerned about thunderstorm potential?" Since the CG lightning summaries used in this study are obviously related to thunderstorm occurrences, this study, with its exhaustive climatological database of indices, should be able to potentially answer their question. In particular, to determine threshold values for individual locations and, if a relationship exists, for a forecast region comprised of the upper-air locations in Figure 1.

Weather forecasters need to know a climatological range of values of each stability index to days with any CG lightning activity. Up until now it appears an exhaustive study of the predictability of the indices to NLDN lightning summaries has yet to be made. A suggested threshold range of values by region was made for thunderstorms by a recent publication from the Air Force Weather Agency (AFWA) titled "Meteorological Techniques" (AFWA TN-98/002, 1998). This AFWA Technical Note suggests a range of values for the indices used in this study and categorizes the range of index values into general thunderstorm, severe thunderstorm, or as tornado indicators (see Table 3). For the purpose of this study, general thunderstorm occurrence or any occurrence for that matter is applicable since it attempts to predict any amount of CG lightning activity. For this study, it should be noted that no inference is made to the severity of each thunderstorm, perhaps for a future study. However, an inference is made as to the potential amount of CG lightning expected later in chapter 5 on regression trees.

Table 3. Suggested range of index values as general thunderstorm indicators (AFWA TN-98/002, 1998).

Index	REGION best applied	Weak (Low)	Moderate	Strong (High risk)
CAPE	East of Rockies	300 to 1000	1000 to 2500	2500 to 5300
K-Index	East of Rockies in moist air	20 to 26	26 to 35	> 35
KO-Index	Cool, moist climates (Pacific	> 6	2 to 6	< 2
Lifted Index	All	0 to 2	-3 to -5	< -5
Showalter	CONUS	> 3	2 to -2	< -3
Total Totals	East of Rockies	44 to 45	46 to 48	> 48
SWEAT (for Severe)	Midwest and Plains	< 275	275 to 300	> 300

The statistical software package SPSS (version 10), allows the user to select a range of values for each index, permitting an easy assessment of the merit of the suggested range of values from Table 3 for each index category (Weak, Moderate, Strong). One must keep in mind that weak, moderate, and strong in this case represents an "indication" for general thunderstorms without reference to their severity.

These threshold ranges for each stability index are used as a starting point to observe any correlations to CG lightning occurrence. Each location's index data was merged effortlessly with the CG lightning data using the statistical software package S-Plus. The merge by variable (MONTH/DAY/YEAR/HOUR) command allowed the dates and times of the indices to sync with the lightning data. An interesting way to indicate the initial relationships between the index values and CG lightning occurrence using SPSS were displayed as simple line plots by month of CG lightning occurrence versus counts of the number of times each index was within the thresholds established in AFWA TN-98/002.

For instance, KO and SSI seem to have an inverse relation. This was a bit confusing at first but realizing that no lower or upper limits were constrained on these two indices suggests that limits should be applied. The SWEAT index matched the best in the summer months, but significantly over-counts in the winter/spring months.

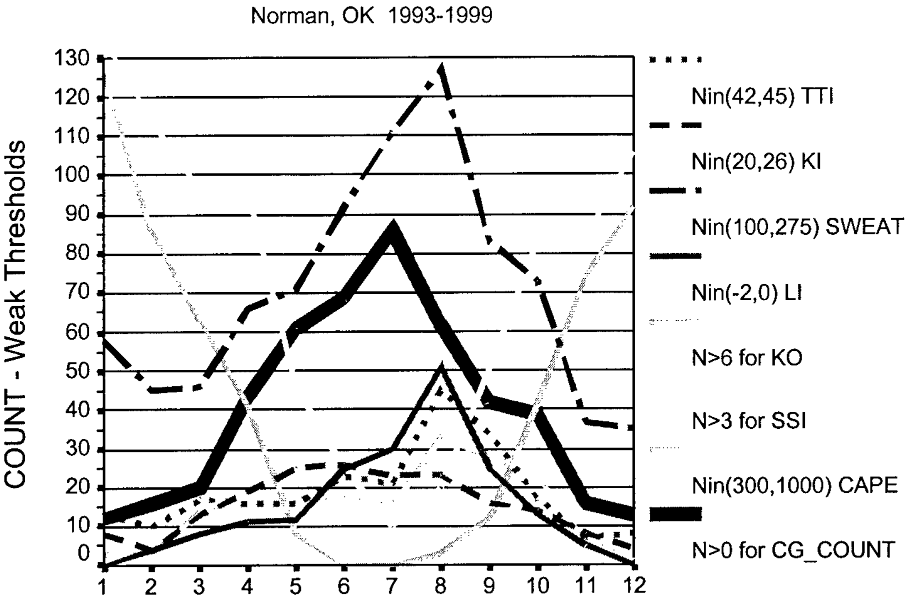


Figure 11. 00Z LBF CG<50nm occurrence in thick line vs. "weak" thresholds for index counts by month.

This suggests different thresholds for the "cool" months might be appropriate by an adjustment of the lower thresholds toward more unstable values. TTI and KI have a reverse relationship in that they seem to grasp a correlation in the "cool" months while dramatically under-counting events during the "warm" active season. In this case, an adjustment should be made to include more unstable values. Before any adjustments were made, comparisons using the "strong" threshold values from AFWA TN-98/002 were compared in Figure 10.

The predictive ability of the indices using the "strong" threshold values indicates that the thresholds established for TTI matched astonishingly well to the occurrence of CG lightning events (N>0 for OBCOUNT). KO thresholds significantly over-counted events while the remaining indices substantially under-counted.

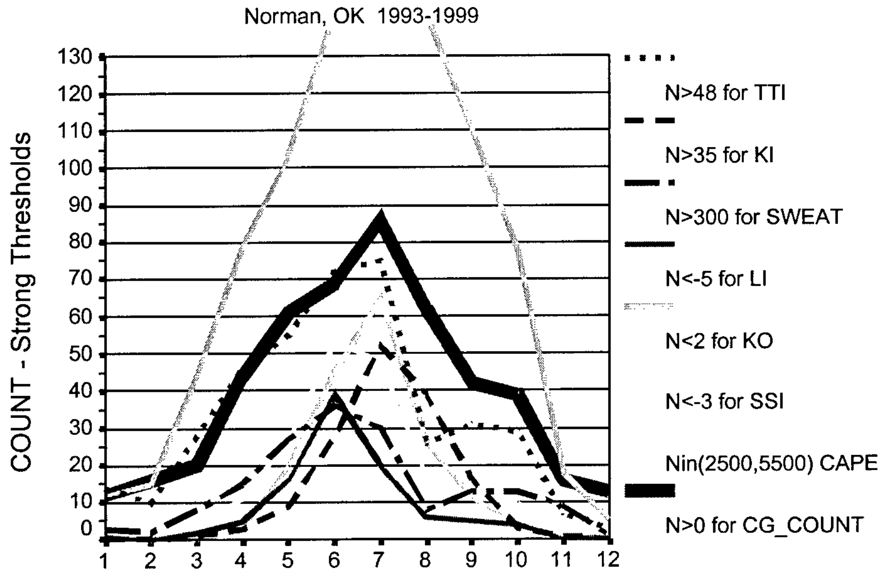


Figure 12. 00Z LBF CG occurrence as thick line vs. "strong" threshold index counts by month.

With these results, a comparison was made and an attempt to determine a more suitable range of threshold values was made while keeping in mind the relationships of the indices found for the "weak" and "strong" thresholds. Results for the best-fit thresholds at LBF and OUN are displayed in Figures 11-14.

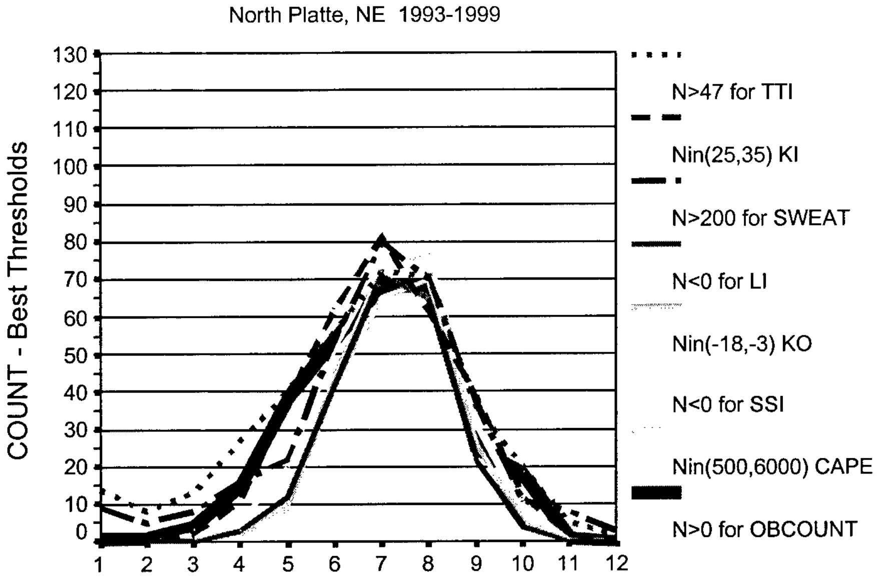


Figure 13. 00Z - LBF Best Annual Index Thresholds.

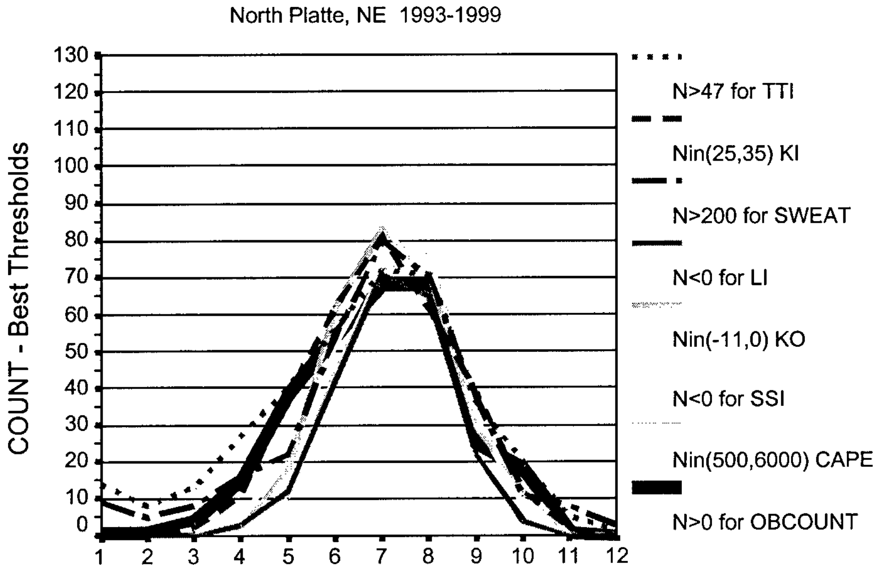


Figure 14. 12Z - LBF Best Annual Index Thresholds.

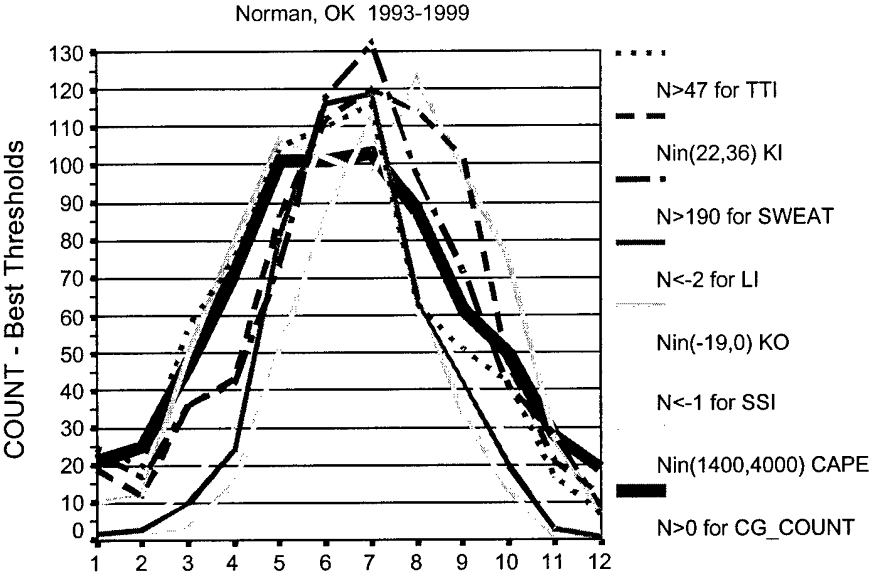


Figure 15. 00Z - OUN Best Annual Index Thresholds.

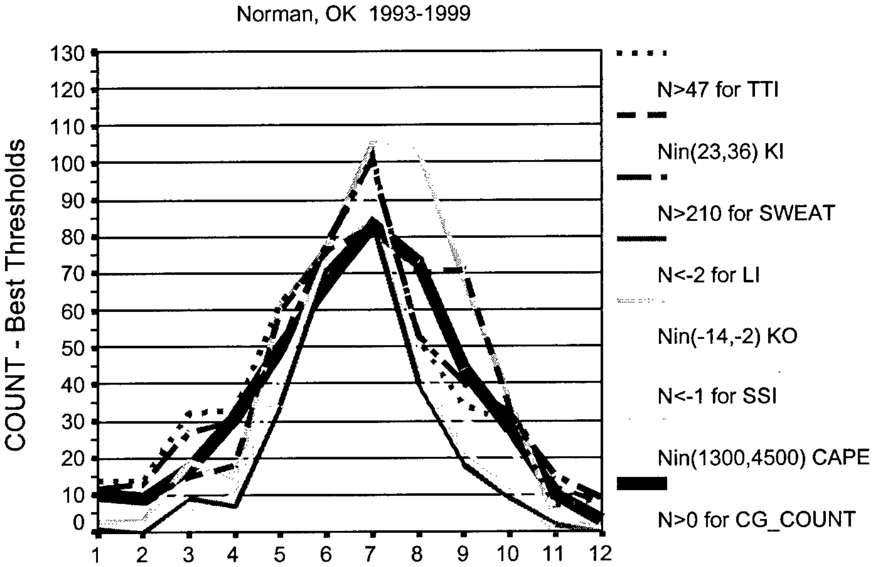


Figure 16. 12Z - OUN Best Annual Index Thresholds.

3.5 A Better Range of Index values

Categorical box and whisker plots of the annual range of index values for 00Z OUN (Figures 15 and 16) help ascertain another way to evaluate the annual range of values each index can take on for days with and without CG lightning (labeled in the figures as <u>none</u> and <u>t-storm</u>). Suggested improvements and results for LBF 00Z and 12Z are displayed in Tables 4 and 5.

A box and whisker diagram illustrates the spread of a set of data about the mean. It also displays the upper quartile, lower quartile and interquartile range of the data with 50% of the data residing inside the "box". A shorter box in this case is indicative of more consistency as a categorical predictor with a narrower range of values.

The annual categorical box and whisker plots once again reveal the decreased variability of most indices for the <u>t-storm</u> category.

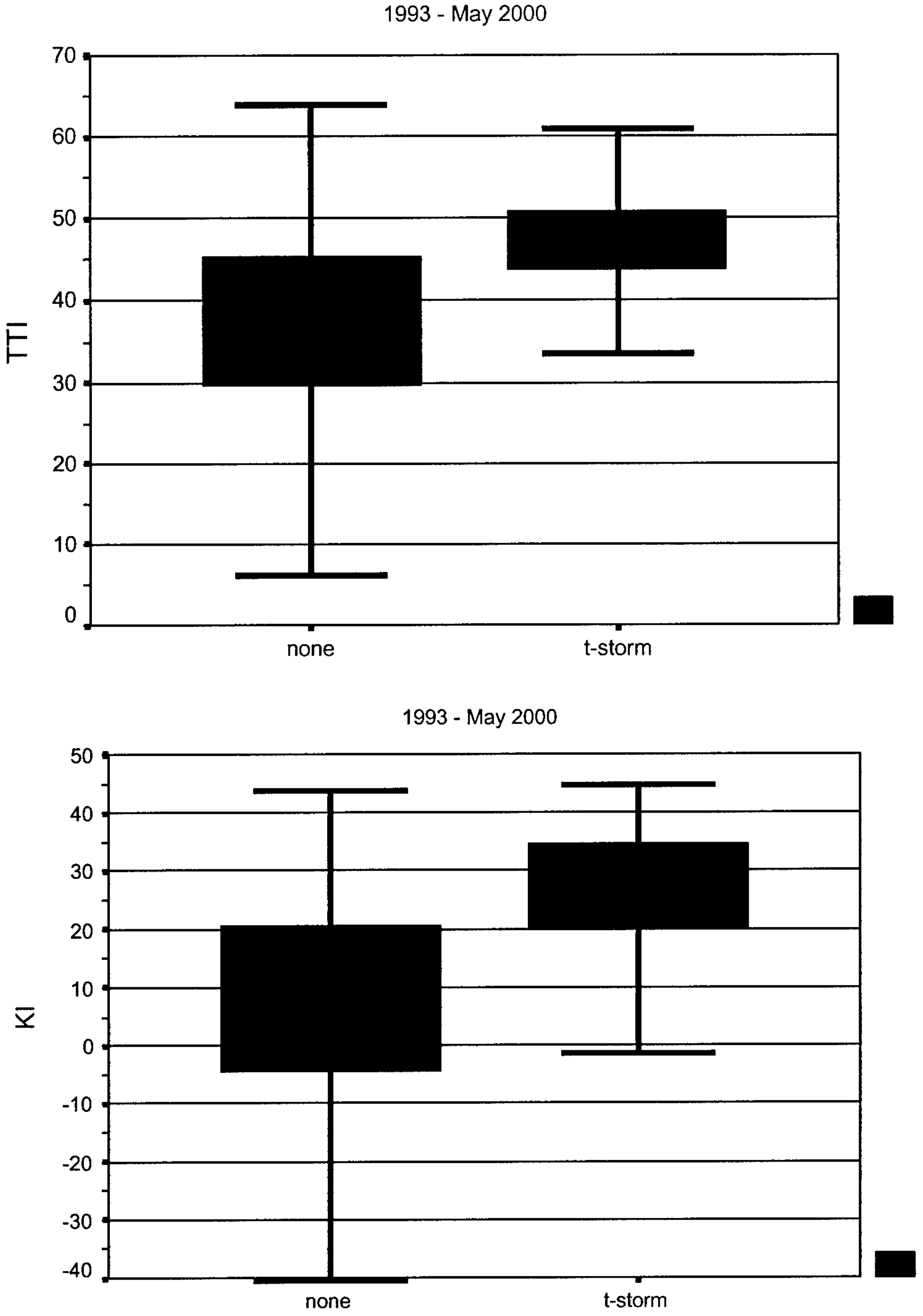


Figure 17. 00Z - OUN categorical box plot of TTI & KI - annual summary.

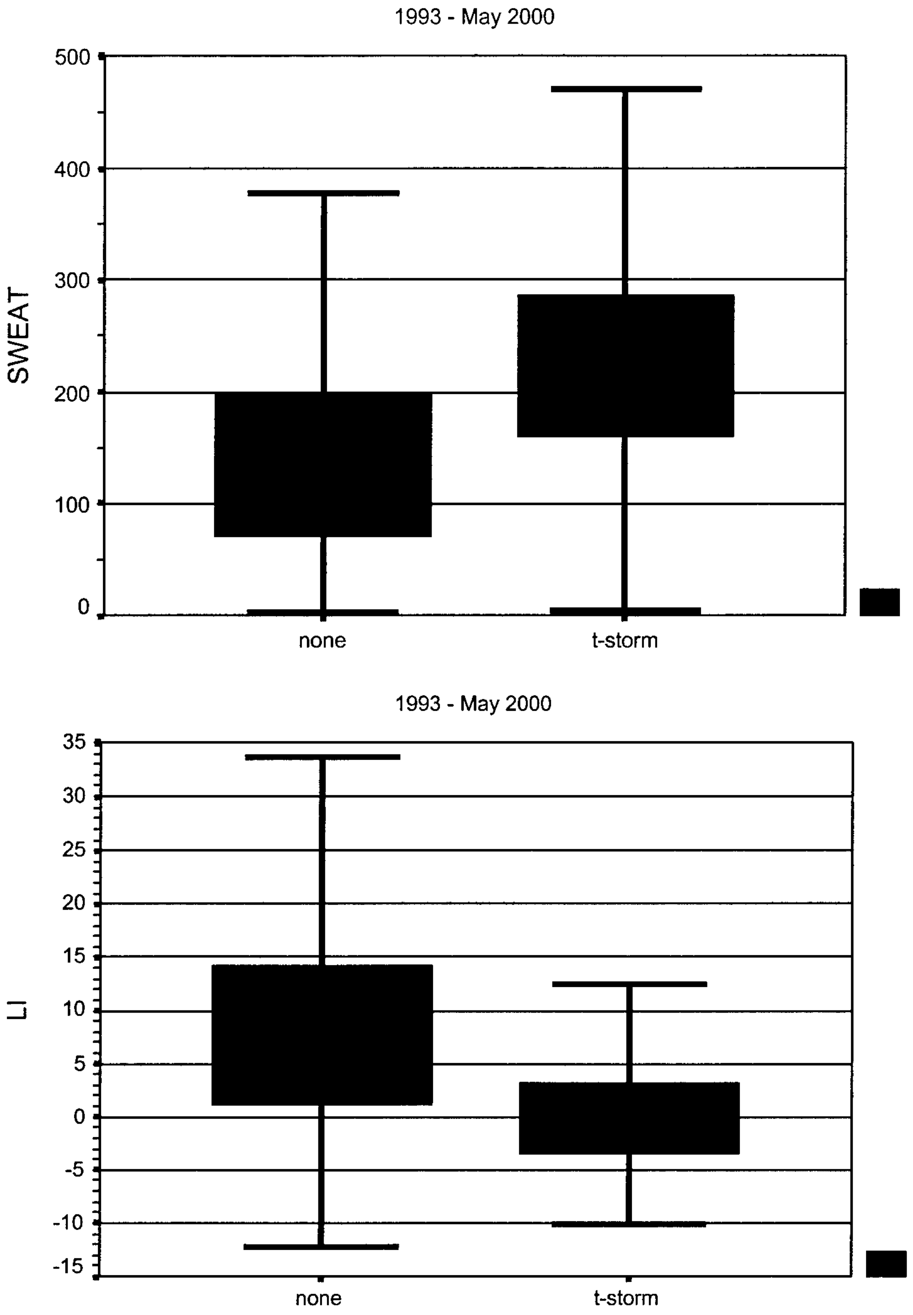


Figure 18. 00Z - OUN categorical box plot of SWEAT & LI - annual summary.

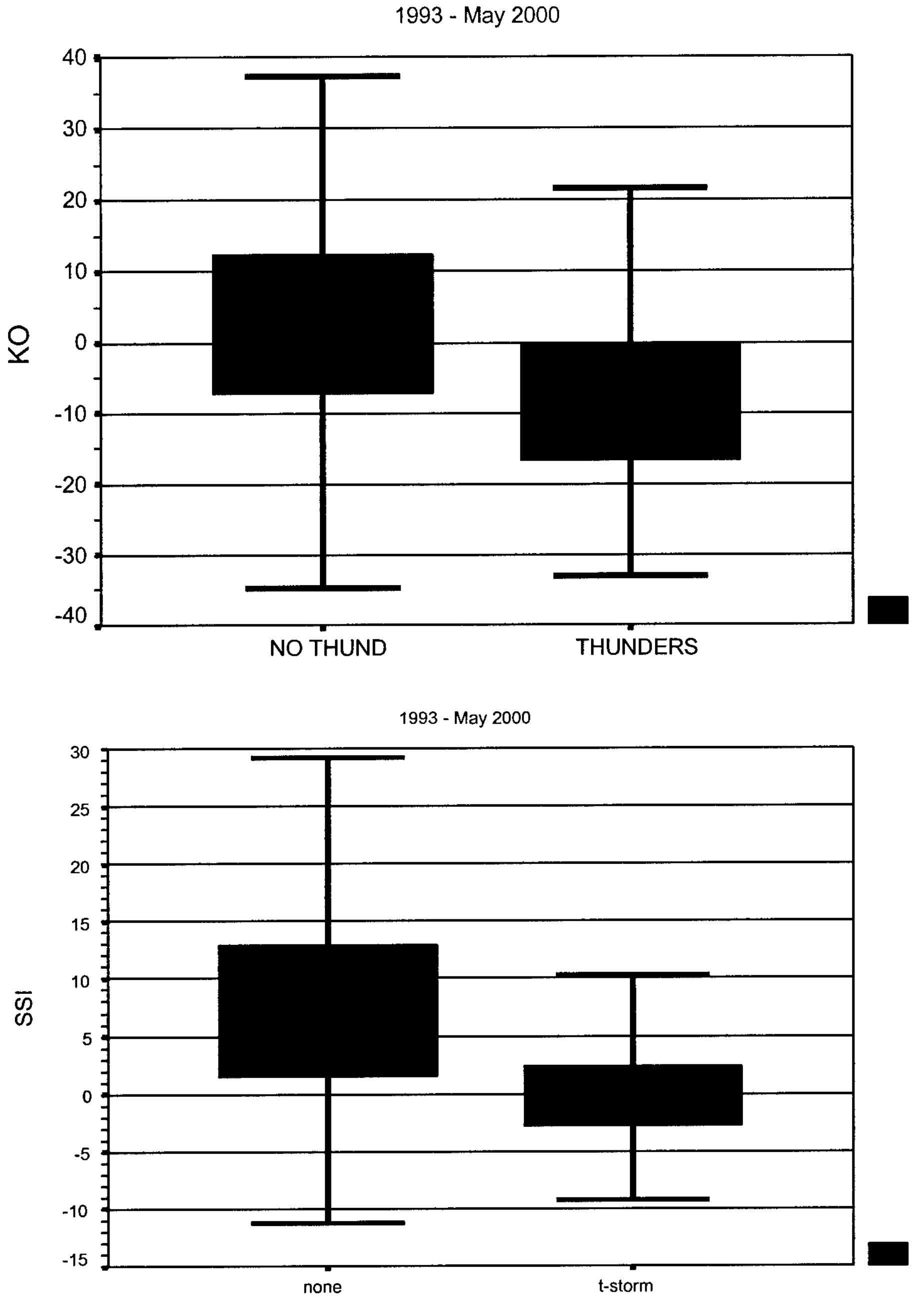


Figure 19. 00Z - OUN categorical box plot of KO & SSI - annual summary.

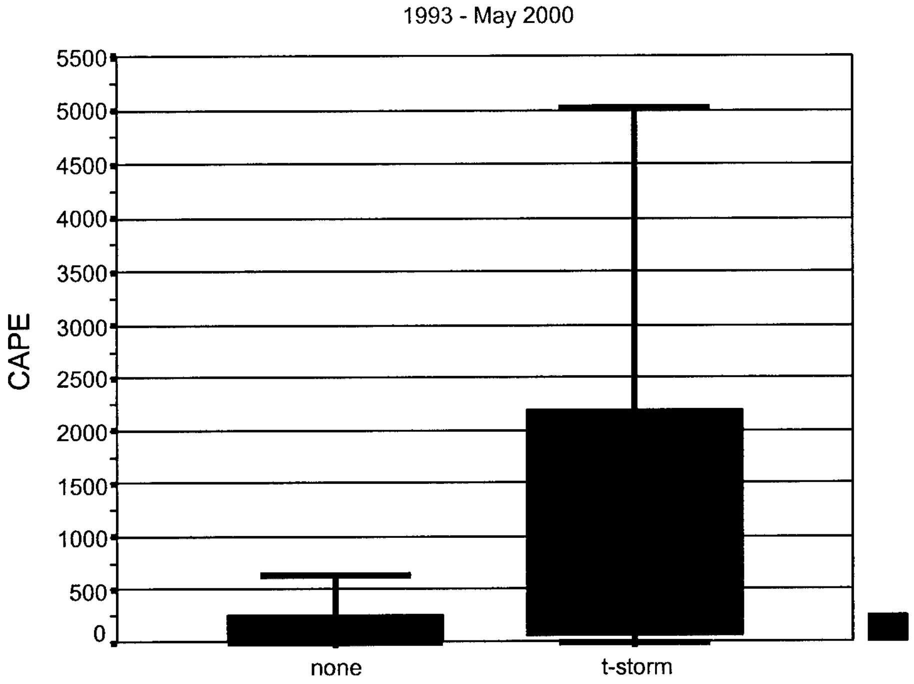


Figure 20. 00Z - OUN categorical box plot of CAPE - annual summary.

With 50% of the range of index values determined inside the "box", there is good agreement as to the hypothesis that the suggested range of values determined in Tables 4 and 5 are superior to the range of values determined for general thunderstorms in AFWA TN-98/002 (Table 3). Indeed, for example, the <u>t-storm</u> "box" for KO in Figure 15 indicates a range of values from -18 to 0 inside the "box", which related well to the suggested threshold range of values for 00Z OUN in the southern plains in Table 5.

Table 4. Suggested range of values for predictive ability of each index to CG lightning occurrence in the Northern Plains.

Index	REGION applied	% time in category	12Z	00Z
CAPE	Northern Plains	66.50%	500 to 6000	500 to 6000
K-Index	Northern Plains	57.10%	25 to 35	25 to 35
KO-Index	Northern Plains	56.30%	(-)11 to 0	(-)18 to -3
Lifted Index	Northern Plains	70.80%	< 0	< 0
Showalter	Northern Plains	71.70%	< 0	< 0
Total Totals	Northern Plains	67.20%	> 47	> 47
SWEAT	Northern Plains	73.90%	> 200	> 200

Table 5. Suggested range of values for predictive ability of each index to CG lightning occurrence in the Southern Plains.

Index	REGION applied	% time in Category	12Z	00Z
CAPE	Southern Plains	56.40%	1300 to 4500	1400 to 4000
K-Index	Southern Plains	49.40%	23 to 36	22 to 36
KO-Index	Southern Plains	48.50%	(-)14 to -2	(-)19 to 0
Lifted Index	Southern Plains	57.30%	< -2	< -2
Showalter	Southern Plains	57.60%	< -1	< -1
Total Totals	Southern Plains	56.50%	> 47	> 47
SWEAT	Southern Plains	51.20%	> 210	> 190

3.6 Summary of Data Analysis Methods

A starting point for assessing the utility of each index to CG lightning within 50nm was made by initially employing the suggested range of index values in AFWA TN-98/002 for general thunderstorms (Table 3). The "weak" threshold range of values seemed to have the lowest relationship and significantly over-counted CG lightning events while the "strong" threshold ranges significantly under-counted them. An improved range of values was determined analytically and suggested in Tables 4 and 5. Wider ranges of values for CAPE were required for 12Z OUN and slightly more unstable values were required for the 00Z sounding thresholds to be more germane. However, no effort was made to imply the severity of each storm event. Instead, these suggested ranges are applicable to any CG lightning event (within 50nm) relative to the indices used. It is assumed in this case that the indices are representative of the atmosphere within a 50nm radius to determine CG lightning occurrence. The suggested range of values found are in agreement with the range in values of the box and whisker plots shown in Tables 4 and 5 for the sampling locations (LBF and OUN).

V. Regression Analysis

In an effort to improve upon the suggested annual range of values of indices best determining CG lightning occurrence established in Chapter 3, regression analyses were conducted to statistically suggest any utility in using individual indices or a combination thereof as predictors of CG lightning.

It is important to note that for regression analysis, only the "active" months 5-9 together are considered for each location using the reasoning established in Chapter 2 on homogeneous datasets. The categorical box and whisker plots in this case should appear less decisive because the non-active months are not considered.

First, linear regression methods were computed with an explanation of the results, limitations, and the obvious disparities with linear regression applications. Next, logistical regression methods were calculated on the occurrence or non-occurrence of CG lightning activity.

In Chapter 6 an in-depth look and introduction to the possibilities of using classification and regression trees as a forecast tool for CG lightning prediction and intensity is explored with motivating results.

5.1 Initial Regression Assessment

Again, categorical box and whisker plots, calculated for months 5-9, were used to contrive the distributions of the predictor variables (indices) as functions of CG lightning occurrence/non-occurrence (labeled as T-storm/none, respectively) in Figures 21, 21, and 23.

Interestingly, comparing the categorical box and whisker plots, most indices appear to have a predictive capability by displaying less variability for CG lightning (T-storm) events and more variability for no (none) events. When predictive ability is considered, least overlap between the none/T-storm categories are desired. No single index stands out significantly, but a few seem less capable or different from the rest. KO and CAPE display the most significant category overlap for both locations, indicating the least predictive capability. SWEAT and CAPE seem different in that they both appear to be the only indices whose variability (length of box) increases noticeably for the T-storm category, while the others are less variable in the T-storm category.

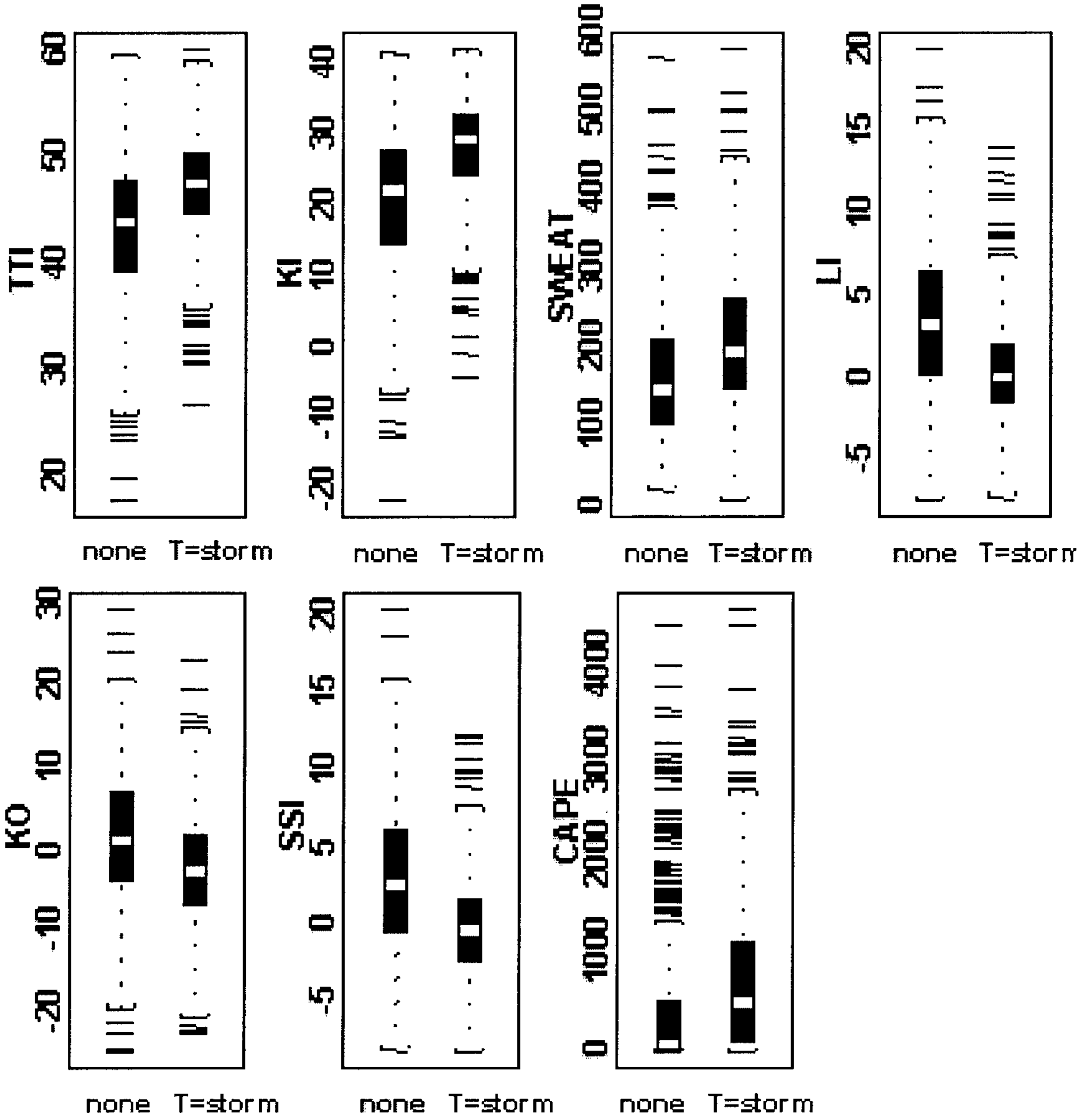


Figure 21. Box and Whisker plots of each index - by category for LBF 12Z (months 5-9).

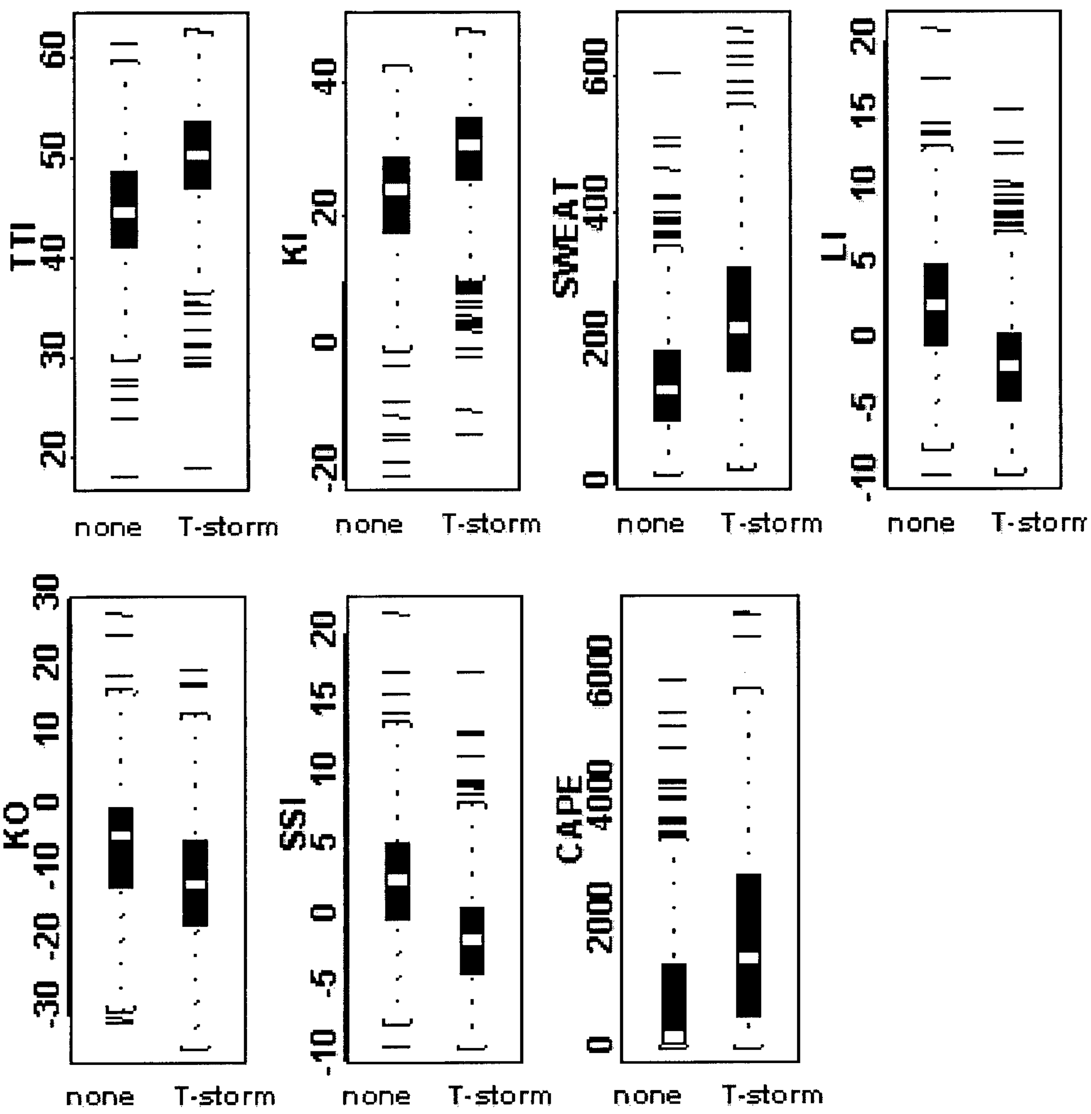


Figure 22. Box and Whisker plots of each index - by category for LBF 00Z (months 5-9).

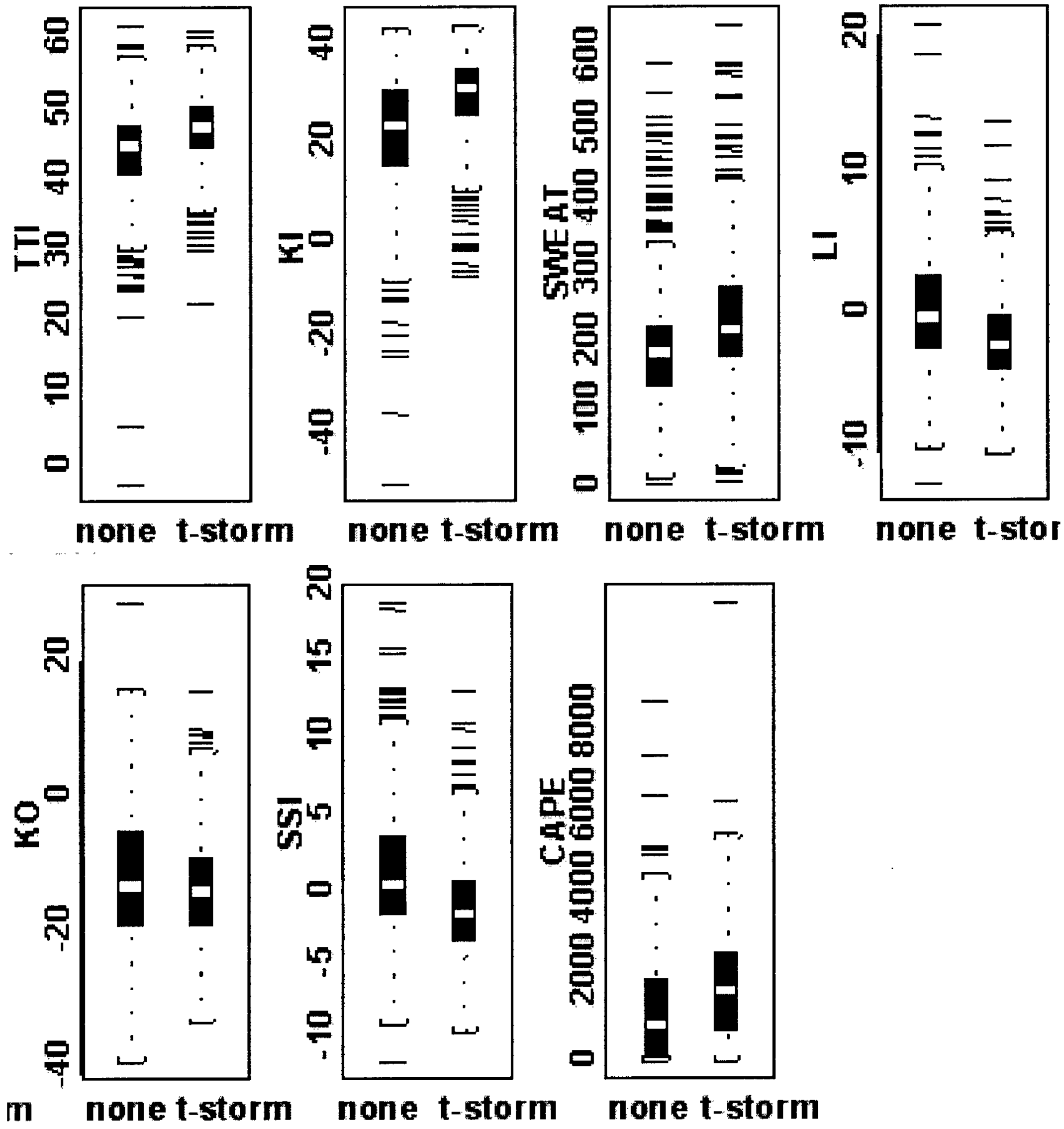


Figure 23. Box and Whisker plots of each index - by category for OUN 00Z (months 5-9).

Comparing 00Z OUN and 00Z LBF - KO, CAPE, and SWEAT appear the least capable predictors. However, at LBF they are somewhat more capable in deciphering between the two categories than at OUN (less category overlap). In fact, KO doesn't appear to be able to distinguish between the two categories at all at OUN in Figure 19. A predictive quality of the indices to CG lightning activity to regression techniques are considered next.

5.2 Stepwise Regression

Stepwise regression is a popular method when searching for good subset models, especially, as in this case, when the number of independent models to compare with is large. Significance was chosen at the 95% confidence level before a variable was considered for model inclusion.

Stepwise regression indicated that SWEAT alone had the most significant relationship. This relationship improved somewhat with the inclusion of TTI. Many of the other variables were dropped from the model due to multiple correlations. R-Squared values ranged from 0.057 at 12Z for OUN to 0.164 at LBF with significance at the 95% confidence levels for the model.

A more detailed linear analysis was computed for 00Z LBF since it showed the highest propensity for a fitted linear model. An improvement of R-Squared values was made by forcing the model equation through the origin so the constant was removed. R-Squared values for fitting a linear regression line to all cases was 0.24 (Figure 20).

Best stepwise linear regression model for CG lightning cases only was:

$$\mathrm{CG}>0 = \mathrm{TTI}*(-6.27973) + \mathrm{SWEAT}*(3.45951) \qquad (7)$$

The model response plot in Figure 21 and 22 show the problem with fitting a linear or even a quadratic regression line to CG lightning activity. A high density of "none" or non-occurrence cases along the x-axis (equation) are observed for a large range of SWEAT values. Figure 22 shows a slight "clean-up" of this density by plotting only the CG lightning occurrence cases. There is still an obvious concentration of scatters at very low CG lightning counts. Perhaps this is evidence that a 100nm radius might be more adequate since more lightning counts would result, but more than likely, the density pattern would remain. Regardless, days without CG lightning (none) have now been eliminated and attention is now focused on linear regression methods to determine CG lightning counts when a CG lightning event is expected. There is less utility under this rationale and the best possible regression fit was through a quadratic expression (see Figure 24).

The quadratic expression is shown as the best R-Squared fit to the regression model using only the SWEAT index. The 95% confidence intervals are superimposed and the R-Squared value is increased to 0.35 - not ideal and only a slightly better fit than when all cases are considered (R-Squared=0.28). The quadratic expression appears to be better at capturing the CG lightning densities at the lower range of values for SWEAT, hence the higher R-Square value.

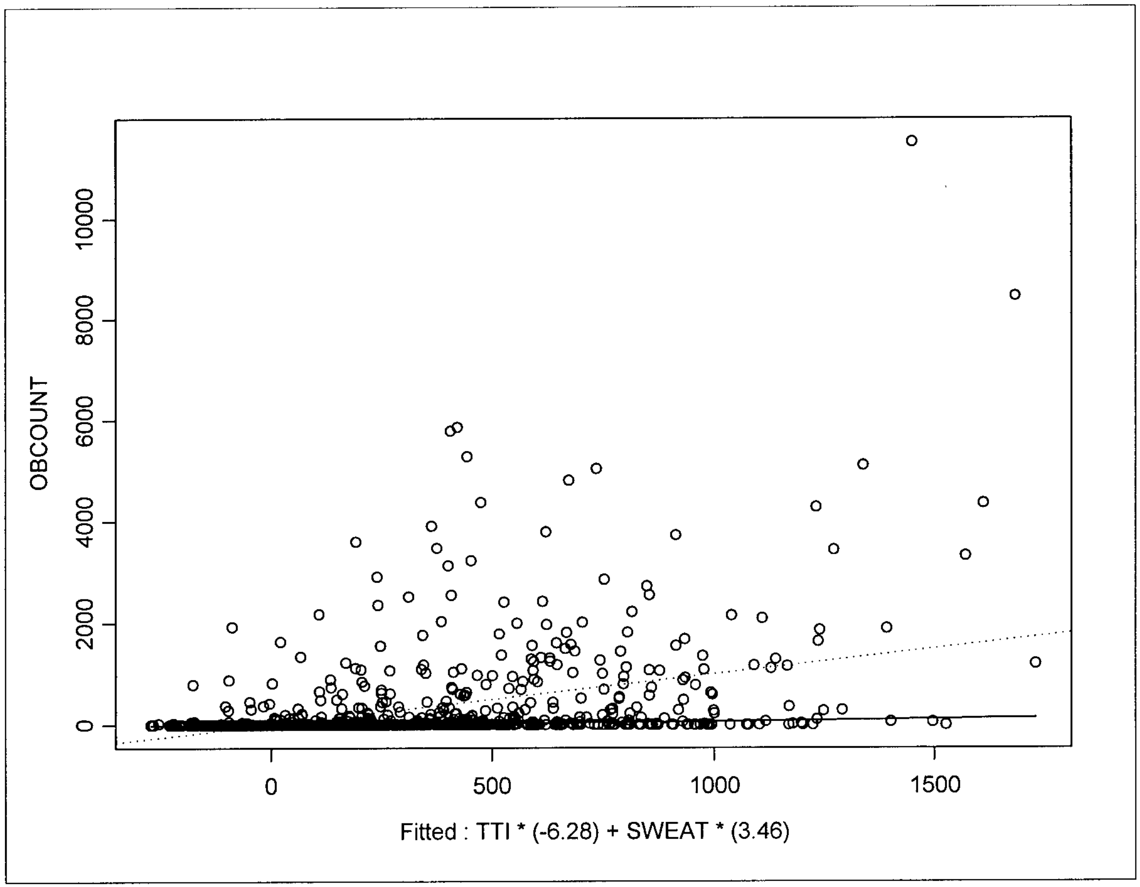


Figure 24. Fitted linear regression results.

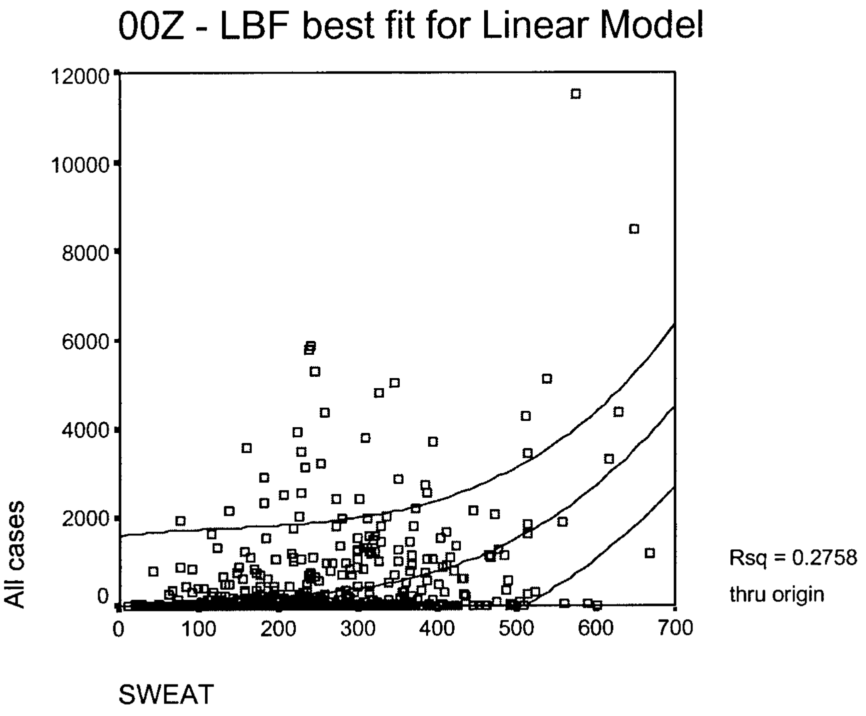


Figure 25. 00Z All cases - LBF best regression fit (QUADRATIC).

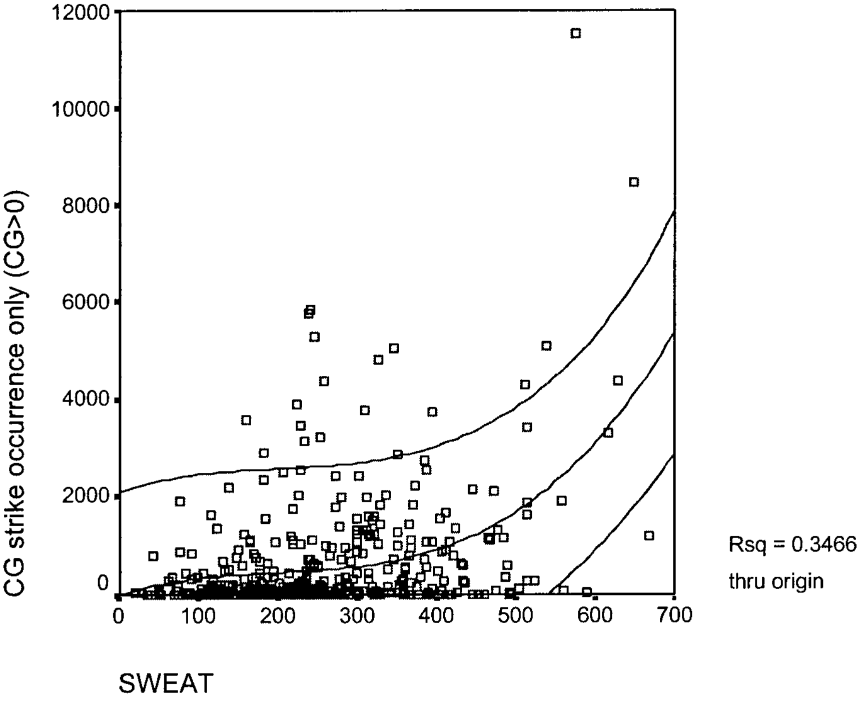


Figure 26. 00Z CG only - LBF best regression fit (QUADRATIC).

5.3 Logistical Regression

Logistical regression analysis extends the techniques of multiple regression analysis to research situations in which the outcome variable is categorical. Logistic regression was used to study how the rate of CG lightning occurrence to non-occurrence depended on the indices as the independent variables. No considerations to CG lightning counts can be made. The interest here was whether CG lightning occurred at all during the valid 12-hour period. A transformation of the data was made in SPSS for each location to add an additional column label "CG.LOG", which stands for CG logistic. A logistic transformation has only two possible outcomes, in this case whether CG lightning did occur (T-STORM) or CG lightning did not occur (NONE). Unlike the linear regression model fit, logistic regression is based on probabilities associated with the values of the categorical predictor (NONE/T-STORM).

The SPSS logistic model results for 00Z OUN with a brief explanation of each test measure are listed in Tables 6 and 7.

The case-processing summary in Table 6 indicates a substantial amount of missing data occurred (30%). This was primarily due to the high missing data rates of SWEAT, KO, and the CAPE indices, which the logistic regression model did not accommodate. Therefore, the model eliminated all cases with any missing values.

The classification table (Table 7) summarizes correct and incorrect estimates of "CG.LOG". The columns are the two predicted values of "CG.LOG" (NONE and T-STORM), while the rows are the two observed (actual) values of "CG.LOG". The overall percentages for both classifications were fairly significant at 75%.

Table 6. Case processing summary.

Unweighted Cases		N	Percent
Selected Cases	Included in Analysis	745	70.0
	Missing Cases	319	30.0
	Total	1064	100.0
Unselected Cases		0	.0
Total		1064	100.0

Table 7. Classification Table.

			Predicted		
			CG.LOG		Percentage Correct
	Observed		NONE	T-STORM	
Step 1	CG.LOG	NONE	276	95	74.4
		T-STORM	91	283	75.7
	Overall Percentage				75.0

The -2 Log likelihood in Table 8 is directly related to the deviance measure used in decision trees in the next chapter and is discussed in greater detail there. A -2 Log likelihood of 794 is rather large and is an indication of the variability of this logistical model fit. The R-Square values are a measure of the strength of association of the indices in the model and their predictive abilities. The association indicated (0.274 and 0.365) has little significance.

Table 8. Model Summary.

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	794.458	.274	.365

The Hosmer and Lemeshow goodness-of-fit test in Table 9 divides the predictors (indices) into deciles based on predicted probabilities, and then computes a chi-square statistic from observed and expected frequencies. The p-value of 0.069 is computed from the chi-square distribution (14.531) with 8 degrees of freedom and indicates that the logistic model has an insignificant fit (Rice, 1994).

Table 9. Hosmer and Lemeshow Test.

Step	Chi-square	df	Sig.
1	14.531	8	.069

VI. Data Mining (DM) and Decision Trees

Traditionally applied statistical methods seem unfocused as a predictive tool due to the enormous variability and range of event versus non-event of the index values. More revealing ways to interrogate the data were sought to possibly improve the results of this study. Originally, it was thought to manually use SPSS utilities to partition a range of values of individual indices and try and find the best probability of event versus non-event of CG lightning. Additionally, this same process was repeated for the CG lightning counts as the response variable to try to establish threshold values of each index, if possible, that best differentiate between active (large number of CG counts) and non-active events. Succeeding at these methods would prove a very useful forecast tool but would require extensive manual work and quickly lead to research for an automated process already developed to handle such a condition. Literature review suggested that the field of Data Mining (DM) might best serve this purpose.

6.1 Data Mining - A Brief History

The DM field was initially born into and developed by the computer realm and was not embraced by the statistical community, initially. Even today there are skeptics, but today it is generally accepted as a useful statistical tool, especially when traditional statistical methods fail.

Data mining is an umbrella term that was initially applied with a negative undertone by the statistical community and the name seemed to stick. Other names applied to DM were "fishing" or "data dredging". It seemed to statisticians of the time that the invalidation of their elegant analytical solutions to inferential problems by exploiting data through "guesswork" had to be errant (Selvin and Stuart, 1966).

The reason decision trees can handle such large databases is their efficiency in computational speed. The concept of DM has largely been a commercial enterprise benefiting computer hardware and software manufacturers that emphasized the high computational abilities associated with DM (Friedman, 1997). Although significant advances in computational speeds over the years have been made, computational speed remains a consideration for new approaches to robust database research. Thus, allowing studies in much larger scope than could be considered before.

There has been an immense amount of research on the uses and applications of DM tools in prediction modeling, the results of which have shown that they can and do surpass the best or normally used models currently in use for some applications. Therefore, DM methods should be taken seriously as a statistical prediction tool.

DM is used to discover patterns and relationships of large observational databases. Statistical software packages such as S-Plus, SPSS, and SAS have recently included DM packages for research professionals to utilize. Some DM techniques include: Decision tree induction, clustering methods, neural and Bayesian networks, and genetic algorithms, to name a few. Decision trees fall under the realm of DM and are therefore introduced as a predictive tool for this study.

6.2 Decision Trees

Decision trees are based on a hierarchal "branched" structure that helps find and plainly display key facets of very large databases. They are important in many DM fields because they are very good at seeing through the "noise" of the data and displaying the most important elements of the results in a straightforward manner (Friedman, 1997). They are hierarchal in that they find the best predictor variables, recursively, and then rank and display them according to their importance in ability to predict the response variable - exactly what was desired in this study. As will be shown, the decision tree method used for this study finds the best index independently (a bivariate response) and is simpler in approach than most other methods. Other methods, such as Oblique Classifier 1 (OC1), allow for a multivariate response regression tree induction system. In other words, OC1 decision trees contain linear combinations of one or more predictors at each tree decision split (Murthy et al., 1994). The result is an oblique split of the data. Oblique splits are said to be more powerful than the simpler univariate test, but also more "expensive" to compute. The term "expensive" has been the benchmark of DM tools in the past because of their efficient algorithms. Especially in its infancy, cost as a measure of computational speed was a much higher priority and selling point for DM.

For the classification and regression trees applied in this study, S-Plus was the program of choice. S-Plus is one of the mainstay programs that incorporate a suite of data mining commands built in - regression trees, K-means

Cluster, and Bayesian Networks to name a few. The difference between other programs with decision tree functions built in, such as SPSS and SAS, is the decision tree algorithm used to determine the best node splits. Improving node split methods for decision trees and other DM tools is ongoing. S-Plus uses reduction in deviance as a measure to find the best discriminator at each tree node.

Many other node split selection techniques exist but past results have shown that no single method is superior to others (Mingers, 1989). Therefore, even though useful and persistent results were found in this study, it is suggested that other tree methods might be considered for comparison. S-Plus tree methods were adapted for this study because they were readily available and assuring results were found when using them. It is proposed that comparisons be made using OC1 and C4.5 decision tree routines that are readily available for research purposes. They are written in S language for operation on Unix platforms (Marmelstein, 1999).

6.3 Applications of Data Mining

Decision trees are one of the main data analysis tools used in DM today (Brodley, C. et al., 1999 and Murthy et al., 1994). Applications of DM tools are being introduced today in many fields. Applications of decision trees in the past are very significant. Some of which include:

Astronomy:

For filtering noise from Hubble Space Telescope (Salzberg et. al., 1995).

Remote Sensing:

For automatic pattern recognition and categorization of earth science data (Rymon, R. and N.M. Short, Jr., 1994).

Hierarchical decision tree classifiers in high-dimensional and large class data (Byungyong, K. and D. Landgrebe, 1991).

Weather Prediction:

Experts in the field of DM are continuously searching for data to exploit. It appears that weather prediction is a relatively new venue to DM and a great potential exists, especially for military weather operations.

The following were just a few examples. Many other real-world applications exist, especially in the bioengineering and medical professions. An interesting example worth mentioning is the application of decision trees for DNA identification by S. Salzberg (1995). In his dissertation, Salzberg applied classification trees to DNA sequences. These sequences involved thousands of base pairs, of which the sequence of interest was the part of the DNA code for proteins that occupied only a small percentage of the sequence. He found that decision trees for this method outperformed any other technique used at the time. His conclusion was that decision trees are "a highly effective tool for identifying protein-coding regions." Regression trees in the past, per se, can find the needle in a haystack, and are highly effective and efficient at it.

6.4 Data Mining in Weather Prediction

DM techniques may be applied in order to generate a more reliable set of decision rules for weather prediction, saving resources and potentially lives (Marmelstein, 1999). A great logical situation exists here. During a preliminary computer study of 328 tornado cases, the Air Weather Service and the National Severe Storms Forecast Center concluded that 14 weather parameters played an important role in the production of severe thunderstorms and tornadoes. This study was conducted prior to 1972 and the parameters chosen were given in order of importance based on computer analysis and forecast experience (Koceilski). The conclusion was that the stability of the atmosphere (easily determined by the indices) was the second most important parameter involved. Some logical questions to ask today would be - "Would this still be true today?" or "Were the datasets used back then comprehensive enough?" There were extreme limitations in data analysis recourses and in the manual techniques applied back then. It would be relatively easy to tap into a much more comprehensive search for important weather parameters through the use of DM tools.

With large databases of weather measurements built up over the years, many useful applications in meteorology may be found by using DM techniques. Decision trees were designed to handle copious amounts of data for quick and efficient calculation and display. More generally, decision trees are basically a series of tests organized in a tree-like structure, where each test on a node split is equivalent to a linear discriminate as in normal regression. In other words, the number of iterations for normal regression is equivalent to the number of nodes in a decision tree. But, unlike normal regression, combinations of nominal/ordinal data may be used as predictors. This is one of the dominant traits of decision trees versus normal statistical regression models. Decision trees have the inherent ability to choose the best predictors among a multivariate set for the given task. As will be seen, the trees grown for this study were quite small because the most significant features were of primary concern. Having a relatively small tree as a forecast tool also makes them both easy to use and to understand. Decision tree experts say one should prefer the simplest model that fits the data (Bishop, 1995).

6.5 S-Plus® Model Used in this Study

The recursive-partitioning algorithm underlying the decision tree function in S-Plus tries to choose the most significant 50/50 split that partition each predictor variable (indices) into increasingly homogeneous regions by a method of reduction in deviance. The result is not only determining the most important index among the others as a predictor for each location, but also the most precise threshold value. To visualize this, imagine a scatter plot of each index divided so that at any node, the split that maximally distinguishes or categorizes the response variable in the left and the right branches is selected. This process is done recursively on each separate predictor variable and determines which index is the single best predictor (using the reduction in deviance goodness of fit measure) for the assigned tree node split.

By applying the reduction in deviance measure, the amount of overlap between the categories (misclassification error rate) is minimized. The average misclassification error rate for most locations ranged between 25-30%, which is similar to the logistic regression classification results. Misclassification error rates suggest that the best results identified are correct 25-30% of the time. This is the best decision tree model fit that can be expected for such large variances seen in the indices.

The tree model used in S-Plus for a classification tree assumes that the response variable follows a multinomial distribution. The multinomial distribution is just a natural extension of the binomial distribution to allow any finite number of categories instead of just two for the binomial.

The two types of decision trees used in this study were classification and regression. Both classification and regression trees were useful tools for predicting CG lightning activity. If the response variable was a factor (categorical), such as t-storm/no t-storm (none), the tree is called a classification tree. If the response variable is numeric, such as CG lightning counts, then a regression tree is calculated.

A summary of the decision tree algorithm process follows these 3 simple steps to determine or "fit" the best results:

1. Split the set of predictors (indices) using the goodness of fit measure (S-Plus uses reduction in deviance). Using the reduction in deviance measure for each potential split in a classification tree is similar to the log-likelihood used in logistic regression for classification trees and poissan/logit regression for regression trees. These tests are done recursively on each predictor-TTI, KI, SWEAT, etc.

2. Check the results of each split comparison. Find the best splits for each index and if every partition is pure, meaning all indices in the partition belong to the same class (none or t-storm), then stop. Label each leaf node with the name of the best class and threshold value.

3. Continue to recursively split any partitions that are not pure.

Figure 27 is an example graphic display of the straightforward manner of S-Plus classification tree output at 00Z OUN. The lengths of the tree branches are proportional to the significance of each classified split, which equates to the quantity of reduction in deviance. Also displayed is a burl plot of each index at the first split, which indicates the goodness of fit summary for each predictor at the model's parent node (KI<27.75). The goodness of fit for each predictor in the model is the difference in deviance between the current node and the successive offspring nodes. The burl plot is a single vertical line for each potential split which is used to determine the best threshold value. Reduction in deviance is plotted against each possible potential split; with the most significant split in this case when the K-Index is

27.75. The worst predictor at this level is by far the KO-Index, noted by the diminutive vertical extent of the burl plot. The KO-Index at the parent node appears to have nearly no prediction capability (reduction in deviance) at any of the potential splits. But the KO-Index is not excluded in any potential future splits. Indeed if we examine the burl plot at the KO<-16.0729 node, in Figure 28, we can see that, although not as decisive as KI was at the parent node, it has the most predictive potential at that level in the tree, compared to the other indices. Further analysis of the data is needed to determine if this split poses any prediction potential.

Figure 29 displays the burl plot at the SWEAT<230.5 node (right side branches of the classification tree). KI and LI also appear to have predictive abilities but the most significant reduction in deviance (highest and steepest peak) is at the SWEAT<230.5 split.

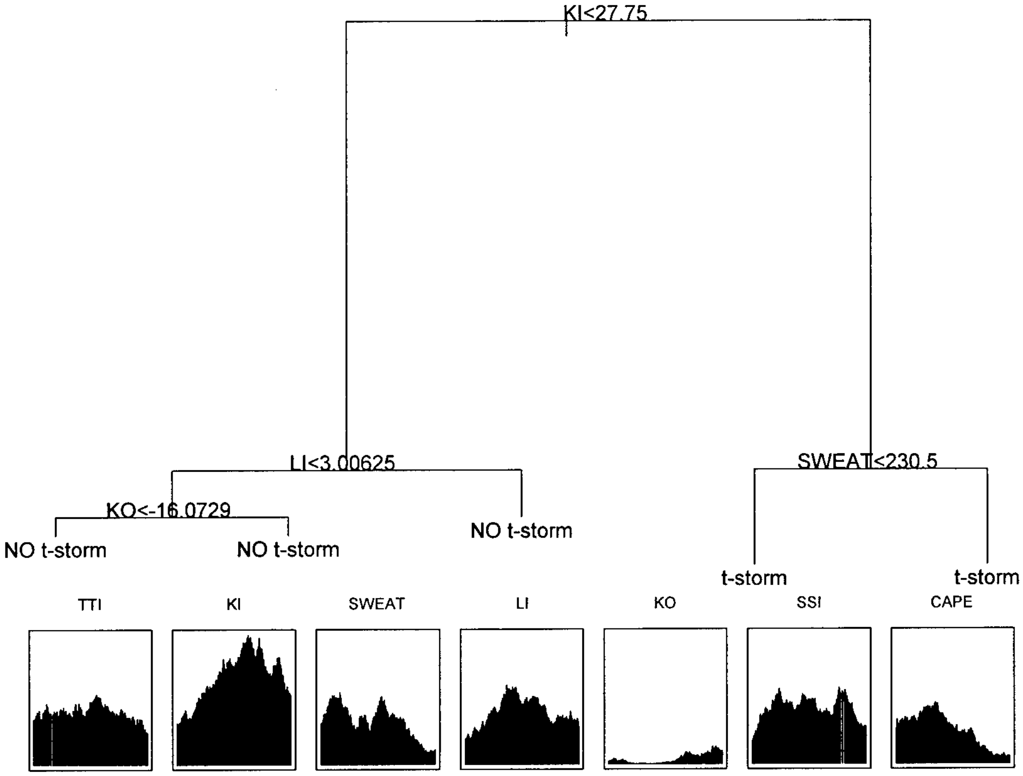


Figure 27. Example classification tree output for 00Z OUN with burl plots of the first tree node split.

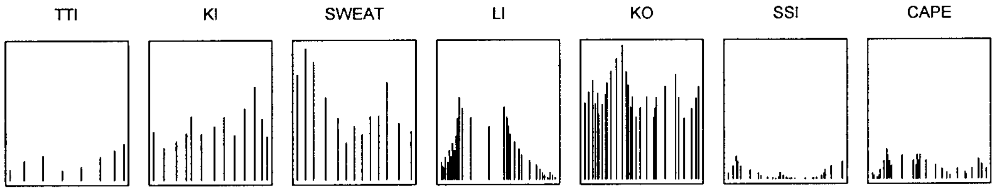


Figure 28. Burl function for 00Z OUN at the KO<-16 tree node split.

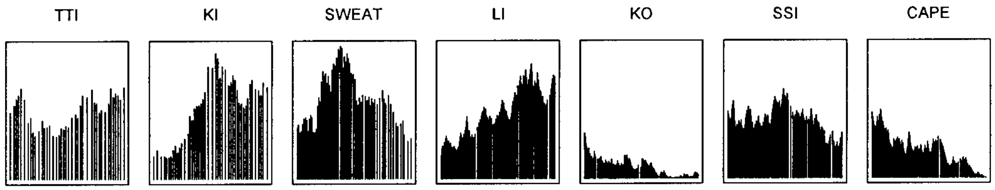


Figure 29. Burl function for 00Z OUN at the SWEAT<230.5 tree node split.

6.10 Classification Tree Summary Output

Actual S-Plus summary tree output in Table 10 is the non-graphic display of the same tree in Figure 27. The significant features are highlighted for simplicity. The classification tree results are not as easily discernable than the graphic, but more detail about what is going on at each non-terminal node branch is possible. An explanation of the summary tree output for 00Z OUN in Table 10 follows: The Parent NODE is split into the first two branches (none and t-storm NODE) which are labeled as nodes 2) and 3) respectively. This first split is the most significant with the significance of all the remaining splits depending on the homogeneity of subsequent index threshold splits, if any. In this example, there are 4 subsequent splits after the first split (5 terminal nodes), their significance depending upon further analysis.

KI<27.75 is the most significant threshold for predicting no CG lightning occurrence (<u>none</u>). There were 335 cases in the <u>none</u> category split with accuracy near 72%. Next, combine this with LI>3.0 in the next branch (node 5) and the probability increases to an 82% occurrence.

KI>27.75 is the most significant threshold for predicting CG lightning occurrence (<u>t-storm</u>). There were

370 cases in the t-storm category split with 60% initial accuracy. When this is combined with SWEAT>230.5 in the next branch (node 7) the probability of CG lightning increases to 72% accuracy.

Table 10. Actual S-Plus summary tree output for 00Z OUN.

```
     *** Tree Model ***
Classification tree:
Number of terminal nodes:  5
Residual mean deviance:  1.234 = 863.8 / 700
Misclassification error rate: 0.33 = 238 / 705
node), split, n, deviance, yval, (yprob)
      * denotes terminal node
Parent NODE:
1) root 705 971.30 Parent NODE ( 0.5461 / 0.4539 )
none NODE:
  2) KI<27.75 335 397.70 none ( 0.7194 0.2806 )
    4) LI<3.00625 229 289.60 none ( 0.6725 0.3275 )
      8) KO<-16.0729 111 129.50 none ( 0.7297 0.2703 ) *
      9) KO>-16.0729 118 156.90 none ( 0.6186 0.3814 ) *
    5) LI>3.00625 106  99.69 none ( 0.8208 0.1792 ) *
t-storm NODE:
  3) KI>27.75 370 494.60 t-storm ( 0.3892 0.6108 )
    6) SWEAT<230.5 195 270.20 t-storm ( 0.4872 0.5128 ) *
    7) SWEAT>230.5 175 207.50 t-storm ( 0.2800 0.7200 ) *
```

As discussed previously, misclassification error rates (or costs) are important for analyzing the significance of classification trees. Misclassification costs become more significant when the categorical counts are low. For this example, these counts can be considered adequate for KI<27.75 (N=335) at the first tree node but at the LI>3.0

threshold in the next tree node, counts are debatable (N=106, which is near the minimum deemed necessary for significance). Note that less than a 15% gain in category prediction between LI<3.0 (node 4) and LI>3.0 (node 5) are revealed (0.67 versus 0.82 respectively), which is not highly significant.

In summary, slight increases in the significance of the results were found in the indices ability to predict no CG lightning events (within a 50nm radius).

6.7 Determining a Significant Decision Tree

At this point it is important to mention that sometimes the significant split determined by the tree may favor one classification over the other. An inherent potential imperfection of tree algorithms results when there are a disproportionate number of classifiers (none/t-storm) or the total number of cases is too small (Fickett, J. and C.S. Tung, 1992). Maximizing the dataset for each location should alleviate this imperfection.

Fickett, J. and C.S. Tung (1992), found that decision trees tend to optimize accuracy on the larger class of data. This appears to be the case for this study as well, especially for some of the 12Z dataset results at some locations due to fewer t-storm category occurrences for the 12-hour period. If annual data was considered, an even more disproportionately higher number of none classifications would result since obviously there are fewer to no classifications of CG strikes in the "cooler" months. For example, the maximized tree models at 12Z for most locations, typically showed the initial (parent node) split of 60/40 (none/t-storm) (see results displayed in Appendix A and B). This may have an influence on the 12Z tree results since optimum initial category split would ideally be 50/50. Consequently, the most significant split, for example, at 12Z LBF was discerned at a questionable threshold, with LI ascertained as the most important index with a threshold value of 4.0. Experience tells us, and comparisons made to nearby locations, suggest that this threshold value is questionable. Also, the normal tendency to split the first node into a decisive category was anomalous in that it chose a higher than normal probability for none but a disproportionately lower probability than normal for t-storm. The first node prediction at 12Z LBF for t-storm was lower than that for none resulting in both first node classifications as none, which is not ideal for the first category split. These facts again support the justification for combining months 5-9, to "equalize" the categories and "clean up" the data.

In Table 11 the summary statistics for each predictor indicates that the best split/threshold values are not just the mean or median. In fact, for 00Z OUN it is actually establishing the split based on the tree model's goodness of fit, which is the best reduction in deviance for S-Plus. The most significant index as a predictor was KI with a threshold value of 27.75 which is somewhere between the median and mean. Also note the significantly higher rates of missing values for SWEAT, KO, and CAPE, which are excluded in the maximized tree models.

Table 11. Summary statistics for 00Z OUN.

```
     Summary (OUN - 00z)

       TTI            KI            SWEAT
Median: 46.40   Median: 28.40  Median:203.5
Mean: 45.77       Mean: 25.55    Mean:218.2
Missing: 37.00 Missing:38.00  Missing:207.0

     KO             SSI             CAPE             LI
Median:-13.170  Median:-0.249  Median:1333.0  Median:-1.3590
Mean: -12.220     Mean: 0.319    Mean:1548.0    Mean:-0.6466
Missing:223.00  Missing:37.00 Missing:236.0  Missing:33.00
```

6.10 The Significance of Missing Data

Missing values can occur either in data used to build trees, or in a set of predictors for which the value of the response variable is to be predicted. There were no missing days for the CG lightning summary data so after merging the two data sets (indices and CG lightning); missing data were found for the indices only. Similar to logistic regression, tree regression permits missing data only in predictor variables. Missing data can be a problem if there are consistent underlying relationships in the reason it was not calculated in the dataset, causing distorted results. For the purpose of generating a decision tree with the highest unambiguous set of classifications, Marmelstein (1999), suggests "filling in" each missing case of the dataset with a fixed value using an imputation method to "repair" them. S-Plus utilizes a built-in feature that enables the user to automatically eliminate the missing attributes or replaces them with a new factor variable; with an added level named "NA" for the missing variables. It then leaves numeric predictors alone even if they contain missing values. Since this study is based on real-world results in a climatological framework, any manipulation or "imputation" method would not be desirable. Instead methods were researched to maximize the database for each location. As a result, using the "full-model" approach might be wavered for an approach that would maximize the usable data for the model as well as be consistent with results.

6.9 Determining Significant Results

Classification (decision) trees have a tendency to purposely over-fit the data. Brieman et al. (1984), whose works on CART are referenced often in data mining literature and is the basis for the development the tree technique used in this study (S-Plus), determined methods for best tree development. For best results, the tree should be over-fitted, in other words, grown too large in order to not miss any key splits that may be hidden. This yields very low to near perfect misclassification error rates but reciprocate reliability. The reasoning behind this is that the data may show a downward trend or insignificant reduction in deviance, and then show a significant trend "hidden" further down the tree. After the tree is grown it should then be "pruned" back for the best fit. Finding the best fit is rather subjective because it depends on the nature of data being used. In this case we are looking at the data sample in more of a real-world climatological sense so interests are in the key splits - the most important ones. We are interested in high probabilities with high occurrences for best model accuracy. Knowing this brings forth the realization that if a node is split significantly above a minimum requirement (say 100 cases) then that node must be highly significant. As a result, most decision tree software have built-in pruning functions to "prune back" the insignificant nodes. One solution to the problem of over-fitting is to reduce or limit the size of the tree in some manner. Over-training can be alleviated in S-Plus by requiring a minimal number of cases before a split is considered. It was found that a minimum of 100 cases worked best to maximize results and minimize over-fitting. Results were maximized in that none of the key variables or node splits were missed or left out because, as will be shown in the output example, more insignificant or less accurate splits are found when sample size splits reach 100 cases. This is supported by the fact that a noticeable number of violations of index stability trends were present for nodes with observations near or below 100. For example, a node further down the 12Z tree modeled for FWD (Fort Worth, TX) with fewer than 100 observations indicated a split that indicated a higher probability for classifying t-storm when KO> -11.9 (see Appendix A). This is normally considered a more stable trend for the KO Index. This is most likely an inconsistent spurious trend because it is likely there were not enough cases involved to show consistent results. Other reasons could be an indication that lower/upper bounds may exist, but in any case, these reasons would not improve the results significantly.

Another way to supplement the proper size of a tree is displayed in Figure 26, which is a reduction in deviance versus number of tree nodes plot. Again, the most significant reduction in deviance is obtained in the first split. Subsequent splits at node 5 indicate the most significant reduction in deviance and results appear minimal thereafter. It was determined that the most significant information obtained from the classification tree results were within the first 5 nodes, with preferred significance given by limiting case counts to 100, which typically resulted within the first 3 nodes.

Marlelstein (1999), Brieman et al. (1994), Mingers (1989), and others, suggest that due to the various methods used in node split selection, for best results it is best to cross-validate the results for consistency. Techniques used to test the "validity" of tree methods can be made by removing a portion of the data before the tree is grown (or trained), then grow the tree on both data sets and compare the results.

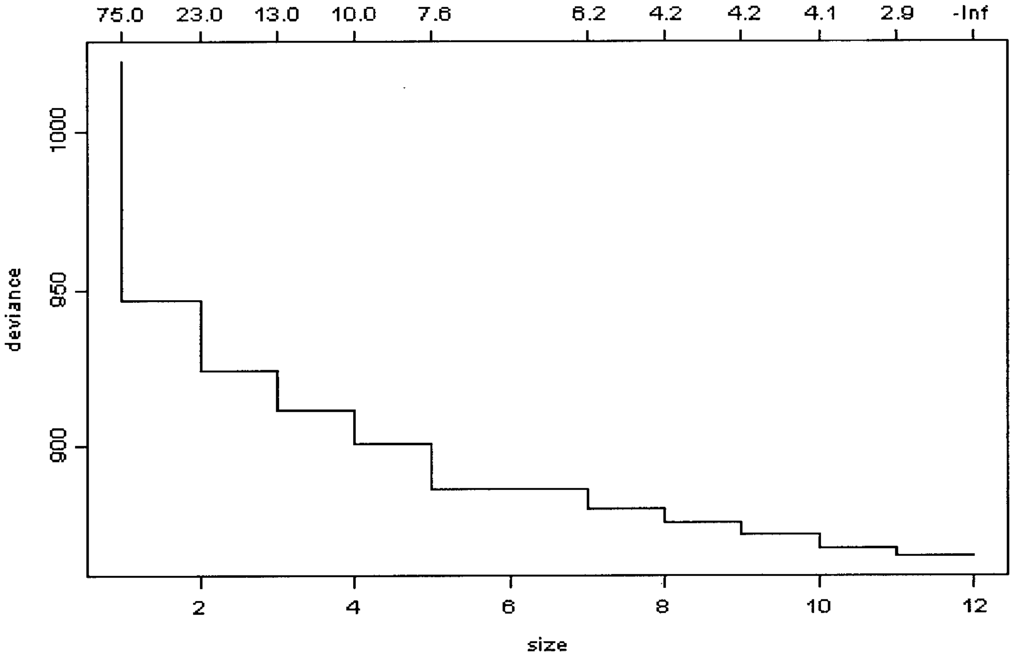


Figure 30. Reduction in deviance versus node size plot for 00Z OUN.

The amount of data to set aside for comparison is highly subjective, but the smaller the test sample, the less likely consistent results will be obtained. Some suggest 10 to 20 percent as a test sample for large databases, while others suggest 40 to 50 percent, if possible. For this study, comparisons of the full model results (SWEAT, KO, and CAPE included) lowered the total case counts for each location by at least 30% due to the missing data of these indices. Consistencies between the split selections of both models for each location indicated very stable results and a high degree of confidence in the outcome. These consistencies are easily determined by comparing the graphical results of both models in Appendix C and D.

For the decision tree used in this study, high probabilities and high occurrences signify very significant results as a forecast tool. So a higher credence should be given to the first split and then only the subsequent splits that indicate a continued large occurrence (N) count along with consistency to surrounding sites. It appears that results with counts near 250-300 cases should be considered highly effective. More confidence can be applied to lower counts nearing 100 cases if the threshold values of the index remain consistent to surrounding locations possessing higher counts. For example, at 00Z OUN (see Appendix B - 00Z Full Model Results - NO T-Storm), KI<10.8 remains a consistent predictor threshold value for NO T-Storm at many surrounding locations: OAX: KI<11.1, TOP: KI<10.6, RAP: KI<10.6, and LZK: KI<11.9, even though the case counts dwindle to near 100 cases, these consistencies have far reaching implications to the significance of the index and associated threshold values determined by the tree model. So KI values near 11.0 indicate a compelling threshold for predicting a non-event nearing 90% accuracy at those locations. In the next section the most significant tree results are compared.

6.10 Decision Tree Results

Appendix A consists of a transformation of the less user-friendly textual S-Plus tree output, as seen in Appendix C, to an easily ascertained forecast tool. References are made to a "maximized" tree model and a "full" model. The full model includes all of the indices in the tree calculations, resulting in data loss caused by missing data in some of the indices (see section on missing data). SWEAT, KO, and CAPE contain the highest missing data rates (over 40% at some locations) so a maximized model was developed that included only KI, SSI, LI, and TTI. The full model is included for cross-validation purposes in Appendix B. The classification tree results may be utilized as a more significant model at some locations for regression tree results due to the importance of SWEAT and CAPE as CG lightning count predictors at those locations.

The tree model results were analyzed for each location, each time period (00Z/12Z), and for both full and maximized models. The most significant results were then quantified in Appendix A. These official summaries are displayed in tabular and graphical form in Appendix A, and present forecasters a user-friendly interface to interpret the index threshold results by geographical region.

Table 12 is a summary of the classification tree results for the maximized model at 00Z OUN and LBF which are tabulated in Appendix A.

Table 12. Sample classification tree results at 00Z from Appendix A for maximized dataset.

	OUN (1009)		
	T-Storm	N	P
if	KI>25.2	605	0.56
&	LI<-1.1	402	0.63
&	KI>35	157	0.74
	No T-Storm	N	P
if	KI<25.2	404	0.75
&	TTI<46.7	293	0.8
&	KI<10.8	107	0.87

	LBF (1037)		
	T-Storm	N	P
if	SSI<1.1	585	0.69
&	KI>30.6	305	0.78
&	TTI>52	147	0.86
	No T-Storm	N	P
if	SSI>1.1	452	0.76
&	SSI>5.6	152	0.84
&	TTI<42.9	122	0.76

Next to the site identifiers (OUN or LBF) in Table 12 are the total number of cases available for calculation by the tree model (in parenthesis). Noticeable decreases in the total number of available cases are seen in the full model results due to missing cases mentioned earlier. Below the location identifiers are the two tree classes determined by the initial split: T-Storm and No T-Storm. These represent the occurrence and non-occurrence of a CG lightning event within 50nm for the valid time period. The first index listed below the T-Storm category gives the most significant index and the threshold value (KI>25.2) for OUN and (SSI<1.1) for LBF. This is the first (parent) split in the tree; therefore the same index will be listed under the No T-Storm category as well. The N and P columns are the number of cases at that tree branch (node) and the probability of the occurrence for that category respectively. Notice that the N cases from the parent split add up to the total number of cases for that location. Each index is listed by importance and is inclusive; such is the hierarchal nature of the classification tree. Inclusive is the reason for the "if", "&", and "or if" statements labeled next to each threshold index. Each combination listed leads to an increased probability of categorical occurrence, but are valid only if each occur inclusive with the other when preceded by an "&" symbol.

The results for OUN in Table 12 should read as follows:

There were a total number of 1009 cases classified. The most significant index and threshold value for classifying T-Storm is when KI>25.2, in which there were 605 cases. At this threshold, 56% of the time CG strikes occurred within 50nm. On the other hand, if KI<25.2 (NO T-Storm category), then, of the 404 cases, no CG lightning strikes occurred 75% of the time. Notice the first split at this location was somewhat offset, favoring the NO T-Storm category. To improve the odds we climb to the next "inclusive" branch in the T-Storm category side of the tree output which suggests if KI>25.2 & LI< -1.1 then 63% of the 402 cases included the occurrence of a CG strike within 50nm. By combining LI< -1.1 with KI>35 (the next tree node), the total number of cases dwindles to 157, but of these 157 cases, 74% of the time there was a CG lightning strike within 50nm. The probability for the No T-Storm category increases to 87% when TTI<46.7 and KI<10.8 occur. This tree split combines a rather stable KI value (<10.8) with TTI<46.7 which only occurred 107 times, but with a significant probability (87%).

6.11 Regional Summary Results

For the maximized tree model at 00Z, KI was the best predictor to CG strike occurrence within 50nm (T-Storm). KI was typically either the most significant or the second most significant by location at 00Z, with threshold values ranging from 25-30 for all locations with probabilities near 70% and high case counts (N). Best results were obtained at LBF, OAX, and DDC, where thresholds of KI>30.5 gave probabilities near 80% with case counts exceeding 300, which were deemed very significant. At OAX, case counts were N=209, but OAX has a lower proportion of total cases (N=800) compared to LBF and DDC (N=1037 and N=1012 respectively), due to sounding data availability problems prior to 1995 at OAX.

Highest probabilities, roughly 75-85%, for CG strike occurrences were obtained when a TTI near 50.0 was combined with other indices at many locations.

It is interesting to note that AMA and RAP were the only locations with LI as the lone significant predictor at 00Z, with AMA requiring a slightly more unstable value (LI< -0.4). The results were fairly significant at these locations with initial probabilities of 75%, increasing to near 90% when combined with other indices (at very unstable threshold values). A hypothesis for this is their High Plains location. Both of these locations reside near or above 1000 feet in elevation (Table 1), which are the highest of all locations used in this study. There seems to be some influence as to the significance of the other indices at these locations. LI is calculated using the average mixing ratio in the lowest 3,000 feet of the sounding. Other indices, like the closely related SSI, strictly use 850mb readings. There is some indication here that the 850mb measurements are inadequate in revealing a relationship between the indices at these elevations. The lowest 3,000 feet method appears more plausible for the higher elevations.

Typically, higher initial probabilities (first tree split) resulted for the NO T-Storm category. This relates to past relationships between weather forecasters and the use of indices. Experience tells a forecaster that stable values of indices indicate a low probability of thunderstorm occurrence with a high degree of confidence. However, unstable values of indices usually signify to a forecaster that further analysis is required. What is revealed here are the significant threshold values to which a forecaster might be able to eliminate the need for a thunderstorm analysis. Initial classification probabilities for NO T-Storm ranged from 75-80% versus 60-70% for T-Storm classifications.

6.12 Regression Tree Results

Comparisons of each model are important for the regression tree results because the number of cases involved is significantly lower for both the maximized and the full model since only cases involving CG lightning strike events were considered. The regression tree results were developed as a forecast tool to help indicate the likelihood of either an active or non-active CG lightning event. It appears that at some locations the CAPE and SWEAT indices are the most significant predictors to the expected “activity” of CG lightning events. The regression tree results tabulated and displayed in Appendix A should be used by forecasters after they first determined via the classification tree results, that there exists a high probability for CG lightning strikes within 50nm for the next 12-hour forecast period.

Table 13 is a summary of the maximized model regression tree results for 00Z OUN and LBF that are tabulated in Appendix A.

Table 13. Sample regression tree results at 00Z from Appendix A for maximized dataset.

mean	N	OUN (437)	N	mean	mean	N	LBF (516)	N	mean
439	334	LI = -4.3	103	1455	409	386	SSI = -4.1	130	1191
227	129	TTI = 45.8	205	572	279	267	SSI = -1.83	119	699

Next to OUN in Table 13, in parenthesis, there were a total of 437 cases of CG lightning strike events for the regression tree to work with. The primary threshold value and index determined was LI= -4.3. It is this threshold value which best deciphers between a more active or less active CG lightning strike event. At the split (LI= -4.3) the mean CG strike count per the N=103 events was 1455 when the LI<= -4.3. It would be confusing to say that the values greater than the threshold index are on the right in this case. The higher mean CG lightning strike counts are on the right corresponding to more unstable index values. In other words, more unstable LI values are more negative. The mean CG lightning strike counts at OUN were 1455 when LI values were less than -4.3 and the mean CG lightning strike counts were 439 for LI values greater than -4.3. This indicates that the mean CG lightning counts were over 300% more active on days when LI is less than -4.3 (439 versus 1455). Another index and threshold value exhibiting potential as a useful predictor was TTI=45.8 and was taken from the next node of the same tree. In this case is it less confusing to say TTI values greater than 45.8 signify a more significant CG lightning strike event since higher TTI values are more unstable. In this case, the mean CG lightning strike counts were over 200% more active (227 versus 572).

It should be noted that the regression tree output in this study indicated significantly high deviance values for all locations. This is to be expected with such a large range in the indices values for active and inactive CG lighting events. The model results may not explain a significant amount of the variability in the model but based on the data presented in this study, it represents the most significant results obtainable through the use of S-Plus decision tree methodology. A student-t test revealed that the means found were statistically different from what would be expected if no relationship existed. The results obtained were also consistent with customary trends and threshold values of the stability indices used and statistically reveal the most significant features for weather forecasters to concentrate on.

More unstable values of each threshold index are required for the regression tree results to best determine an active event. Perhaps these values may also correlate well to severe storms outbreaks. This is something that is left for future research.

VII. Conclusions and Recommendations

7.1 Conclusions

This study reveals the feasibility of using atmospheric stability indices to forecast the occurrence of CG lightning activity for the "active" lightning months of May through September (Objective 1). This study's approach was empirical in nature and represents the likelihood of CG lighting probabilities based on past occurrences. The study first suggests an improved range of threshold values, on an annual basis, than those provided in the past for general thunderstorm occurrences. These should be implemented for the stability indices when predicting CG lightning activity, which is closely related to thunderstorm occurrence (Objective 2). The Midwest upper-air stations studied were divided into northern and southern regions and a slight modification for the annual threshold ranges was required for a few of the specific indices, depending upon their location and sounding observation time. The utility of these thresholds was most useful in the northern Midwest where the most constructive indices were the LI, SSI, TTI, SWEAT, and CAPE. CG lightning occurred between 67-74% of the time when these indices were within the determined thresholds. The KI and KO indices had a 56-57% accuracy, the utility of which is questionable.

Alternatively, the annual threshold ranges determined for the southern plains region of the study barely exceeded 50% accuracy for any of the indices. This leaves the threshold ranges found to barely possess any predictive ability at all, based on the threshold ranges established. This region is more active in the winter months and as a result the false-alarm rate for this region is much greater. The influence of CG lightning events in the winter months is much stronger for the southern Midwest. In fact, many of the locations in the northern Midwest had little to no CG lightning events during the winter. Box and whisker plots revealed that the indices were much more variable in the winter months as well.

It was determined that due to the seasonal variations of the indices, especially for the southern Midwest region, the active months (5-9) should be examined exclusively for further study. It should be noted then that the results for the rest of this study are for the combined active months (5-9).

Linear and non-linear regression techniques were applied next to examine the CG lightning data and stability indices for any predictive relationships (Objective 3) that would improve upon the threshold ranges determined earlier. Stepwise linear regression eliminated all but a few specific indices for the best model fit, but even then no significant relationships were found.

Since traditional statistical methods failed to find any significant relationships, new methods of predicting CG lightning activity using stability indices were explored using decision trees from new data mining techniques (Objective 4). Reliable and significant results were obtained and a new predictive forecasting tool was developed that allows weather forecasters to predict the occurrence or non-occurrence of CG lighting events with an average probability of between 80-90% (Objective 5). The most relevant indices and threshold values were determined for each individual location and sounding times. Decision trees implemented an inclusive or hierarchal classification approach while at the same time maximizing the inclusive event counts. This inclusive approach means that observing one index threshold, under the condition that other indices thresholds must occur as well, allows for the significant probabilities found.

Interestingly, the most significant indices and threshold values determined for each location by the decision tree lead to a predictable sequence. The Lifted Index was determined best for use in the high plains locations (RAP and AMA) for both sounding times, in part due to their higher station elevations. Next, the Showalter Index was most significant for the northern plains region of the study at 00Z. Further south in the more moist regions of the Midwest, the K-Index was the most significant at 00Z, most likely due to the fact that the K-Index provides an extra measurement for moisture at the 700mb height level. This extra measurement at 700mb was also proven to be significant for the 12Z sounding times as well since the K-Index significance was predominant for most locations at 12Z. This was easily explained by the fact that the morning temperature inversions commonly found at 12Z during the active months (5-9) in the Midwest could not be resolved by the 850mb temperature/moisture measurements determined by most of the other indices.

The classification tree results developed allow forecasters to determine the probability of a CG lightning event and, if a forecaster determines that CG lightning is expected, the regression tree results allow weather forecasters to determine the potential frequency or "amount" of the CG lightning activity that is to be expected. These results were then displayed in a user-friendly format by location and time in both graphical and tabular forms as a forecast tool for users (Appendix A is written as a ready to use forecast tool for users by displaying these results). Regression tree results displayed the most significant stability index and threshold value for each location whose value above/below gave a 300-500% increase/decrease in mean CG lighting activity based on each threshold found. Again, only events where CG lightning did occur were analyzed under the regression trees since the classification tree results where first used to determine if an event was expected (Appendix A is written as a ready to use forecast tool for users by displaying these regression tree results in both graphical and tabular forms).

The ability of a weather forecaster to predict the probability of the occurrence or non-occurrence of CG lightning for all locations analyzed generally exceeded the 80-90% levels which has far reaching implications. Additionally, using stability indices to determine the expected amount of CG lightning is unique. Therefore, the results of this study should prove to be a useful forecast tool in the operational environment.

7.2 Recommendations for Future Study

Other techniques of analyzing the datasets used during the course of this study were discovered that could ultimately improve upon the results, but time constraints prohibited their implementation in this study. A suggested approach is to develop forecast stability indices generated operational forecast models and compare them to CG lightning activity in the same manner employed in this study.

Another approach is to implement a specialized predictor to the indices. One type of specialized predictor is sometimes referred to as an interactive predictor. Interactive predictors are especially important when forecasting rare events such as severe thunderstorms and tornados. One example of an interactive predictor used to forecast thunderstorms is the KF predictor (Reap and Foster, 1979), which is the KI multiplied by the thunderstorm relative frequency. This predictor forces the climatology (the relative frequencies) to be more responsive to the current synoptic situation. In other words, it applies a weighting factor empirically, based on the past history of CG lightning strike probabilities.

Similarly, over 6 years (93-00) of CG lightning data are utilized in this study and could be implemented to create monthly frequency distributions of CG lightning strikes (within 50nm). These monthly frequency distributions might be useful as an additional input to regression analyses (Reap and Foster, 1979). Also, since this study demonstrated that decision tree analysis revealed more promising results than regression analysis, Table 14 suggests an example method to be used in the same manner decision trees were employed in this study.

Table 14. Example modification of indices that could be offered as predictors to the screening classification/regression tree analyses.

KI multiplied by CG lightning relative frequency.
SWEAT index multiplied by CG lightning relative frequency.
TTI multiplied by CG lightning relative frequency.
LI multiplied by CG lightning relative frequency.

7.3 *Future Data Mining Applications*

There are many suggestions for the use of data mining tools in weather research since it appears data mining techniques are in their infancy in this field. Of relevance to this study though are ways to utilize the stability indices as predictors for the occurrence of CG lightning strikes and the potential number of CG strikes that may be received.

A careful computerized/technical review of the most important forecasting parameters, as summarized by Miller (1972) and as developed by the Air Weather Service and National Severe Storms Forecast Center (Koceilski), could easily be revalidated with the use of data mining methods. Suggested weak, moderate and strong thresholds were suggested in the study but new, more significant, thresholds could still be discovered. Classification trees might, in fact, be capable of determining the single most important threshold value and predictor to focus a weather forecast analysis on, assuming the database used is large enough for an empirical approach. Other additional parameters could also be considered as well. Consistent results among data mining tools indicate to weather forecasters what weather parameters they should concentrate their analyses on. Benefits may also include substantial analysis timesavings as well as increased forecast accuracy.

7.4 *Other Atmospheric Stability Indices to Consider*

It would be ideal to assess the potential of all available atmospheric stability indices, but algorithm development and time constraints were prohibitive for this study. Some of the indices not included in this study but which are suggested for future study as predictors of CG lightning activity are:

- the Fawbush-Miller Stability Index (FMI)
- the Martin Index (MI)
- the Modified Lifted Index (MLI)
- the Bulk Richardson Number (R)
- the Dynamic Index, and the
- Wet-Bulb Zero (WBZ) Height Index

For the purpose of this study, the one index that was not available but which would have been a significant consideration for future research is the wet-bulb zero (WBZ) height because of its recent utility in lightning research. Traditionally WBZ heights are used to forecast hail since certain threshold value ranges correlate well with large hail events at the surface (Miller et al., 1972). Miller showed that a large majority of the reported surface hail occurred when WBZ heights are between 5,000-

12,000ft above ground level (AGL) while large hail is most likely when WBZ heights are between 7,000-11,000ft AGL. Again, restrictions to these values as well as any other atmospheric stability index exist by location and forecast regime and should be determined for individual locations.

7.5 *Development of a Lightning Index*

Finding an improved range of values for hail occurrence by location would be useful, but in relation to this study, another application to consider is WBZ heights and its recent application to the study of lightning occurrence. Theory on the origin of lightning suggests that the process of collision and coalescence of frozen particles in thunderstorms is the primary mechanism for the charge separation that produces lightning in thunderstorms (Dye, 1990). The development of a new “Lightning Index” is currently ongoing. Stuart et al. (1998), suggests a "Lightning Index" would likely be based on specific thresholds of meteorological parameters, such as stability indices, and offer some form of prediction capability for the production and frequency of lightning on a daily basis. Additionally, his suggestion of CAPE and LI to indicate the potential strength of updrafts and instability potential when combined with WBZ heights should be studied. This would provide information that may indicate the potential for the production of frozen particles, which is thought to be important to the formation of lightning based on the theory of Dye (1990).

Stuart et al. (1998), suggests the use of CAPE and LI, but decision tree results from this study suggest the significance of SSI in the northern region of the study, KI in the southern region, and LI in the high western plains region as the most significant predictors to the occurrence of CG lightning events. Perhaps the development of a "Lightning Index" should consider the significant indices found in the results of this study for their use as predictors instead, since geographic location is considered as well. The results of this study also suggest more unstable threshold values of the indices are required when applying to the frequency (or amount) of CG lighting expected.

7.6 *Implementation of Results*

The results of this study using classification and regression trees were significant enough to implement immediately as a forecast tool for the operational weather forecast environment. Appendix A of this study is written as a "ready-to-use" forecast tool for weather forecasters. It is suggested that Air Force Weather units in the Midwest U.S. use this "innovative" forecast tool immediately for forecasting CG lightning activity.

Appendix A: Optimal Decision Tree Maximized Model Results

This appendix is written as a stand-alone forecast tool taken from the thesis research results of Capt. Ken Venzke, Air Force Institute of Technology, Wright-Patterson AFB, OH. It summarizes the official decision tree results to assist forecasters in determining the probability of lightning activity or non-activity for individual upper-air sounding locations in the Midwest U.S. This forecast tool is valid for the "active" months of May to September. The stability indices determined as the most significant by this study were the Showalter (SSI), K-Index (KI), Total Totals (TTI), and Lifted Index (LI).

First, a brief description is made on how to interpolate the results, followed by the official results in graphical and tabular form for both 00Z and 12Z valid sounding times.

To begin, an example tabular summary is referenced in Table A-1 along with the same summary in graphic form in Figures A-1 and A-2. The two upper-air sounding locations are OAX (Omaha, NE) and TOP (Topeka, KS). The number in parenthesis next to the locations is the total number of observations surveyed. It should be noted that only the active months May to September from 1993 to 2000 were assessed for this study. There are two categories derived for the probability (**P**) of the number (**N**) of occurrence/non-occurrences of lightning events (T-Storm/No T-storm). The results are also inclusive, which is the reason for the "if", "&", and a few "or if" statements labeled next to each stability index threshold. This inclusive approach means that observing one index threshold, under the condition that other index thresholds must occur as well, allows for the significant probabilities found. Each combination listed leads to an increased probability of categorical occurrence, but are valid only if each occur inclusively of the initial index threshold value.

The example for OAX in Table A-1 should read as follows: There were a total of 800 observations available from 1993 to 2000. The most significant stability index at this location was when SSI<1.1, of which, 66% of the time a thunderstorm occurred within 50nm of the station during the valid 12 hour sounding time period. This probability increased to 78% when both SSI<1.3 and KI>30.5 occurred. Finally, the maximum probability found for a T-Storm event at OAX was 87% with the additional requirement that TTI>50.1 must occur in combination with the other two thresholds. Alternatively, the maximum probability (91%) found for a non-event (No T-Storm) is when the combinations SSI>1.3, LI>2.5, and KI<11.1 occur inclusively.

Table A-1. 00Z Tabular summary classification example.

	OAX (800)		
	T-Storm	N	P
if	SSI<1.3	377	0.66
&	KI>30.5	209	0.78
&	TTI>50.1	101	0.87
	No T-Storm	N	P
if	SSI>1.3	423	0.73
&	LI>2.5	296	0.81
&	KI<11.1	106	0.91

	TOP (930)		
	T-Storm	N	P
if	SSI<2.2	550	0.63
&	KI>22.9	441	0.69
&	TTI>49	120	0.74
&	KI>35.2	130	0.82
	No T-Storm	N	P
if	SSI>2.2	380	0.76
&	KI<10.6	124	0.9

Figure A-1. 00Z Graphical classification results example for T-Storm probability.

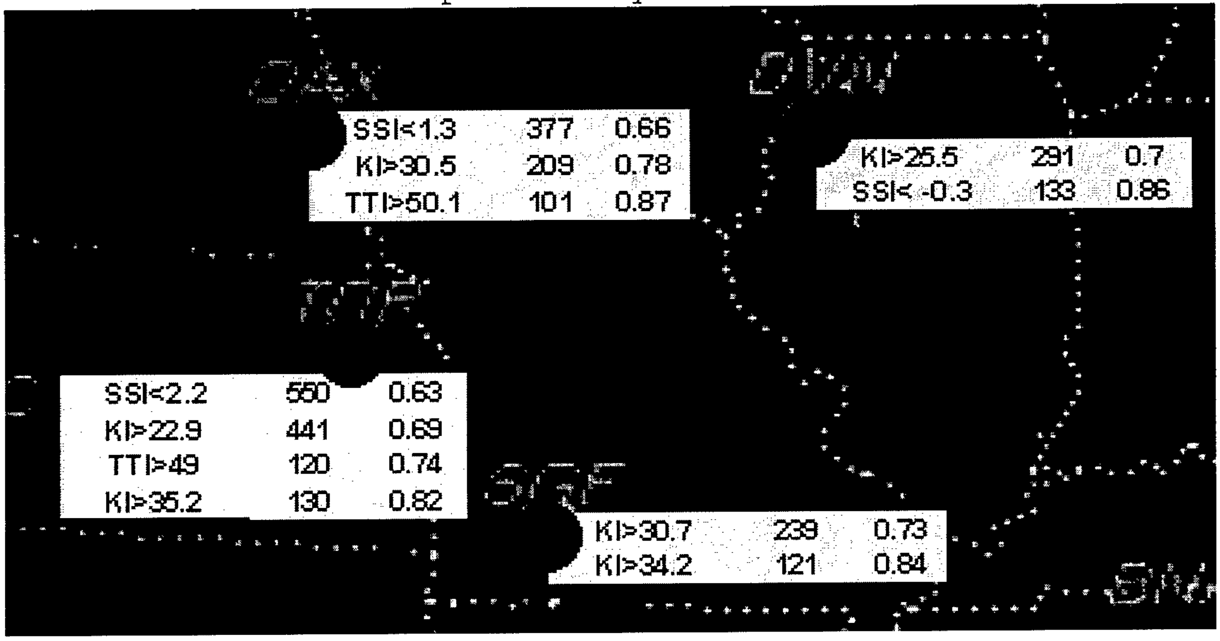
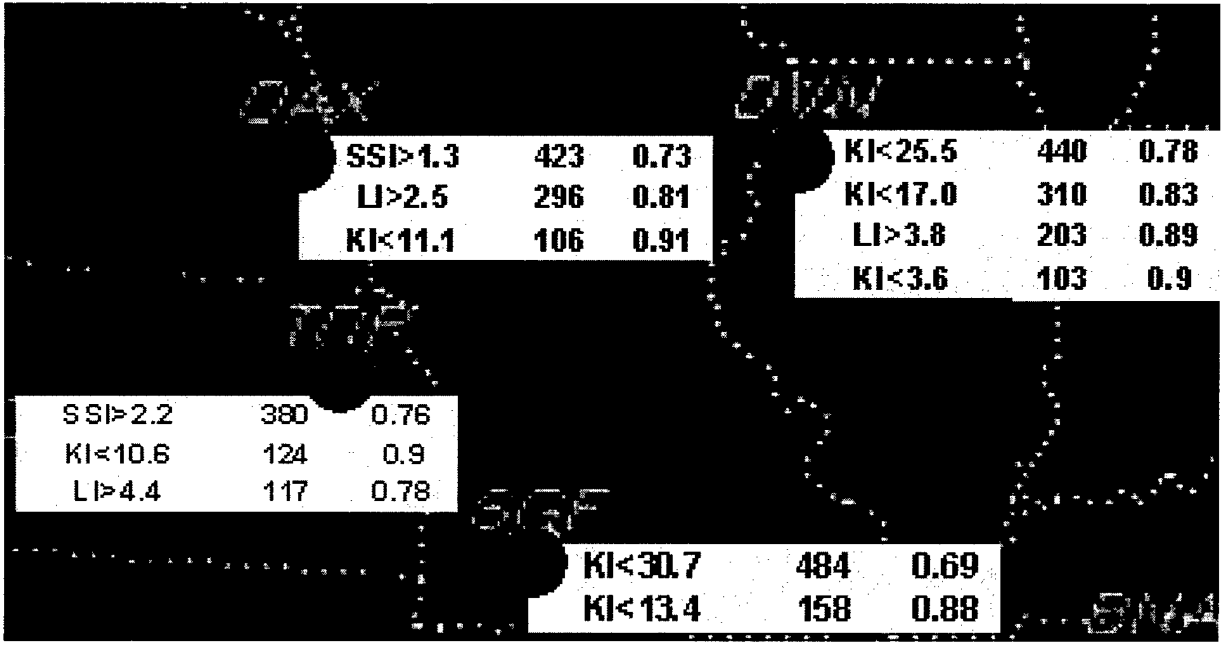


Figure A-2. 00Z Maximized classification tree example for No T-Storm probability.



SSI>1.3	423	0.73
LI>2.5	296	0.81
KI<11.1	106	0.91

KI<25.5	440	0.78
KI<17.0	310	0.83
LI>3.8	203	0.89
KI<3.6	103	0.9

SSI>2.2	380	0.76
KI<10.6	124	0.9
LI>4.4	117	0.78

KI<30.7	484	0.69
KI<13.4	158	0.88

The official graphic results (Figures A-3 and A-4) allow the forecaster to "visualize" the results geographically. In summary, the highest probabilities, roughly 75-85% for a lightning event, were obtained when a TTI>50.0 was combined with other indices at many locations, but the most significant index (which is always listed first) and threshold value, that must occur first, varied by location. Interestingly, the indices and threshold values determined for each location lead to a predictable sequence. The LI was determined best for use in the high plains locations (RAP and AMA) for both sounding times, in part due to their higher station elevations. Next, the SSI was most significant for the northern Midwest region of the study at 00Z. Further south, in the more moist regions of the Midwest, the KI was the most significant at 00Z, most likely due to the fact that the KI provides an extra measurement for moisture at the 700mb height level. This extra measurement at 700mb was also proven to be significant for the 12Z sounding times as well since the KI significance was predominant for most locations at 12Z. This was easily explained by the fact that the morning temperature inversions commonly found at 12Z during the active months (May to September) in the Midwest could not be resolved by the 850mb temperature/moisture measurements via the other indices.

The mean strike threshold results are displayed in graphical form in Figures A-7 and A-8, and in tabular form in Tables A-4 and A-5. The classification results developed allow forecasters to determine the probability of a lightning event and, if a forecaster determines that lightning is expected, the mean strike results allow weather forecasters to determine the potential frequency or "amount" of lightning activity that is to be expected.

The mean strike results displayed the most significant stability indices and threshold values for each location whose value above/below gave a 300-500% increase/decrease in the mean lighting activity for each event. Only events where lightning did occur were analyzed under the mean strike results since the classification results where first used to determine if an event was expected.

Next to OUN in Table A-2, in parenthesis, there were a total of 437 cases of lightning strike events for the mean strike results to work with.

Table A-2. Sample mean strike results at 00Z for OUN and LBF.

mean	N	OUN (437)	N	mean	mean	N	LBF (516)	N	mean
439	334	LI = -4.3	103	1455	409	386	SSI = -4.1	130	1191
227	129	TTI = 45.8	205	572	279	267	SSI = -1.83	119	699

The primary threshold value and index determined was LI= -4.3. This threshold value best deciphers between a more active or less active lightning strike event. At the split (LI= -4.3), the mean lightning strike count for the N=103 events was 1455 when the LI<= -4.3. It would be confusing to say that the values greater than the threshold index are on the right in this case. The higher mean lightning strike counts are on the right corresponding to more unstable stability index values. In other words, more unstable LI values are more negative in this case. The mean lightning strike counts at OUN were 1455 when LI values were less than -4.3 and the mean lightning strike counts were 439 for LI values greater than -4.3. This indicates that the mean lightning strike counts were over 300% more active on days when LI is less than -4.3 (439 versus 1455). Another index and threshold value exhibiting potential as a useful predictor was TTI = 45.8 and was also considered significant for this location. In this case is it less confusing to say TTI values greater than 45.8 signify a more active lightning strike event since higher TTI values are more unstable. In this case, the mean lightning strike counts were over 200% more active (227 versus 572).

The ability of a weather forecaster to predict the probability of the occurrence or non-occurrence of lightning for all locations analyzed generally exceeded the 80-90% probability levels, which has far reaching implications. Additionally, using stability indices to determine the expected amount of lightning strike counts is unique. The results of this study should prove to be a useful forecast tool in the operational environment.

Figure A-3. 00Z Lightning probability results.

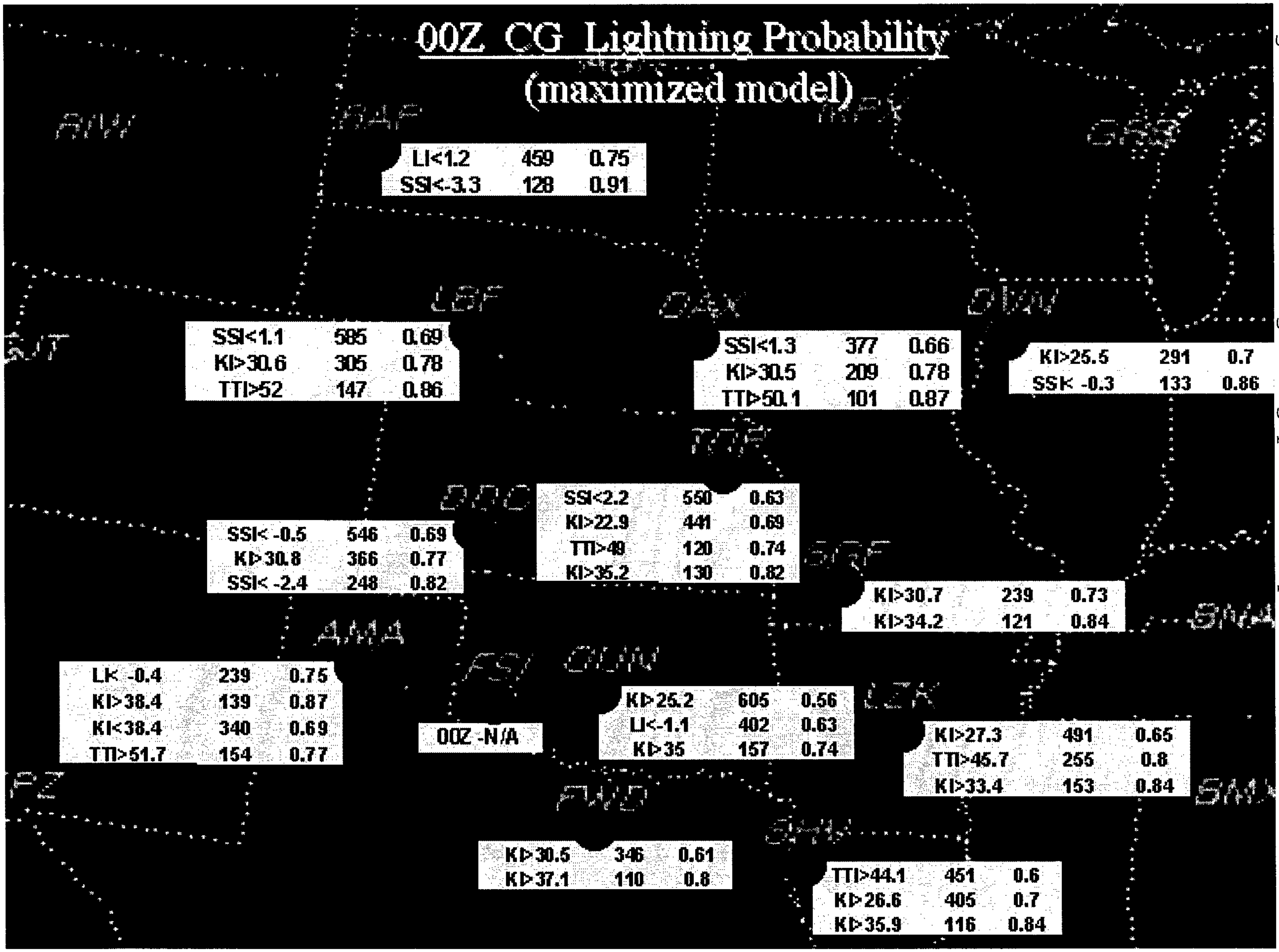


Figure A-4. 12Z Lightning probability results.

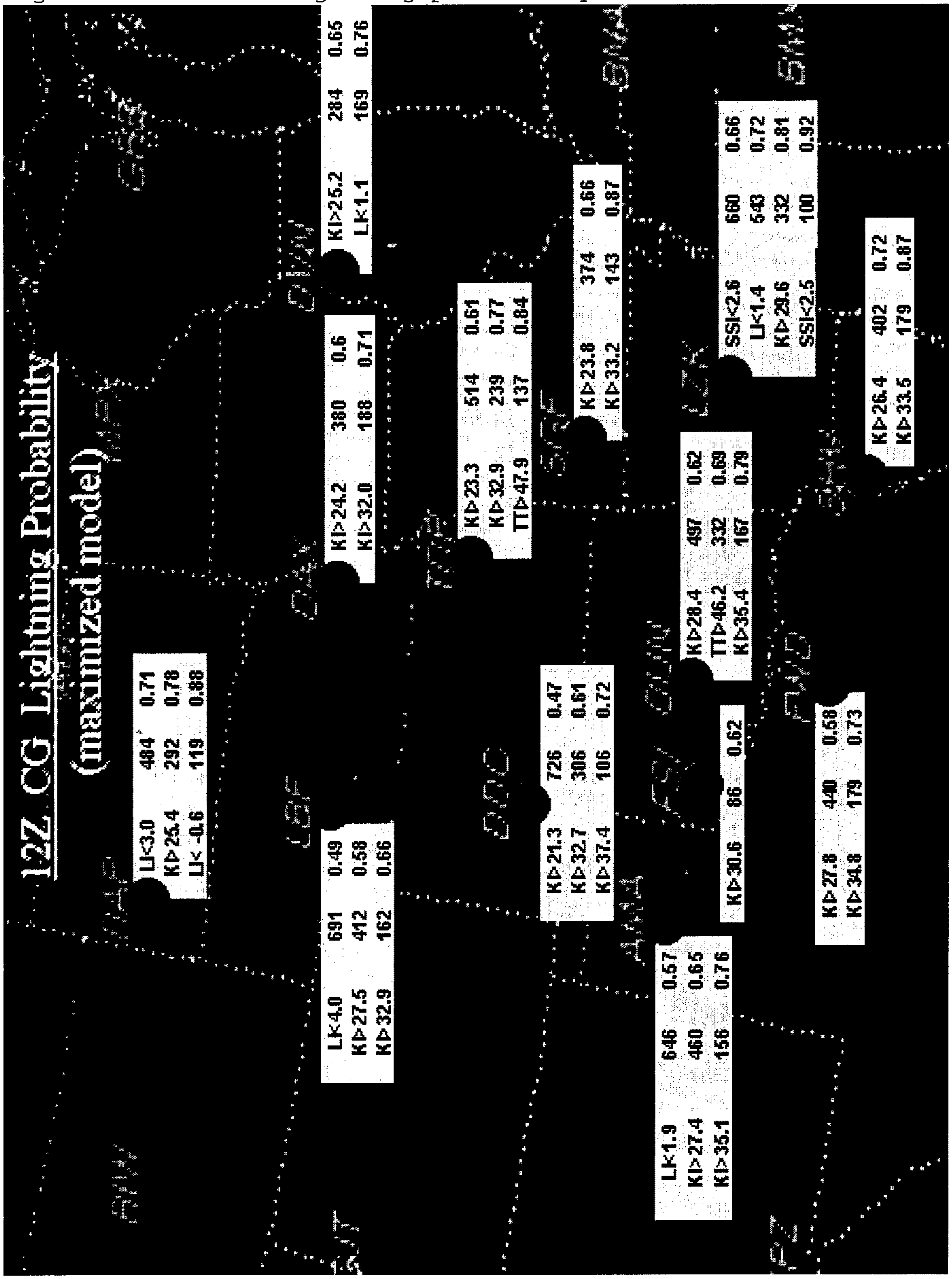


Figure A-5.   00Z NO Lightning probability results.

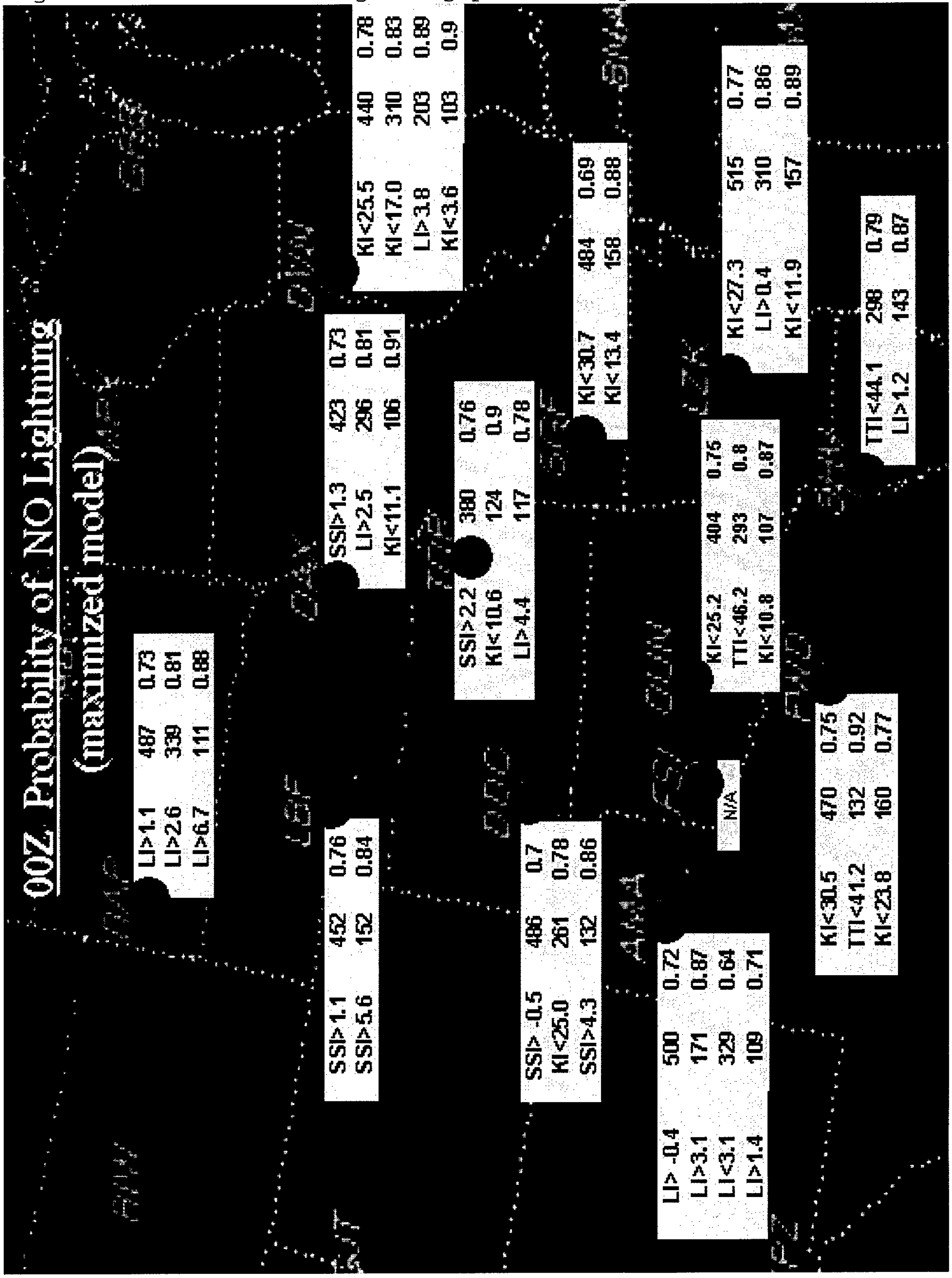


Figure A-6. 12Z NO Lightning probability results.

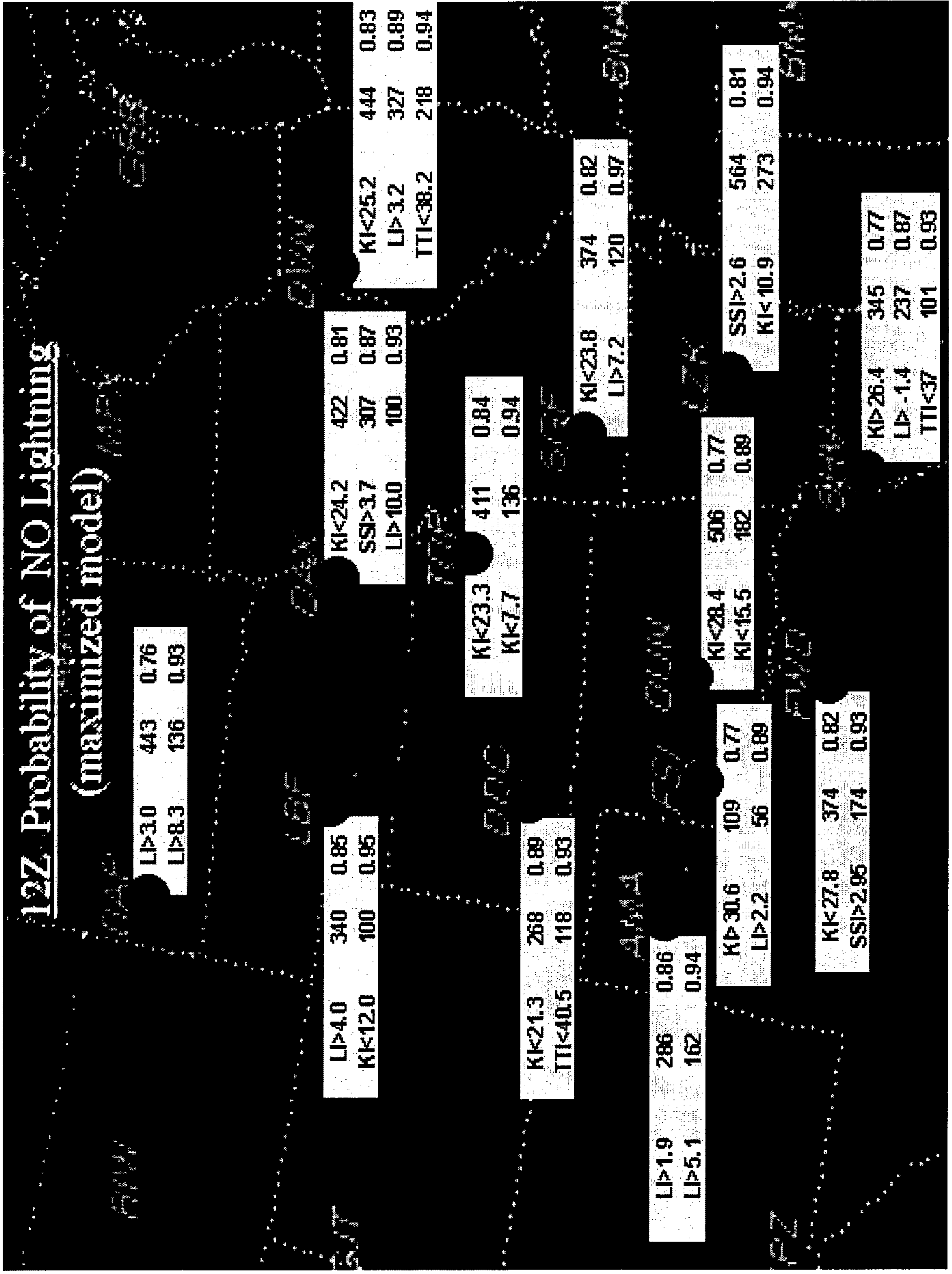


Figure A-7. 00Z mean strike thresholds results.

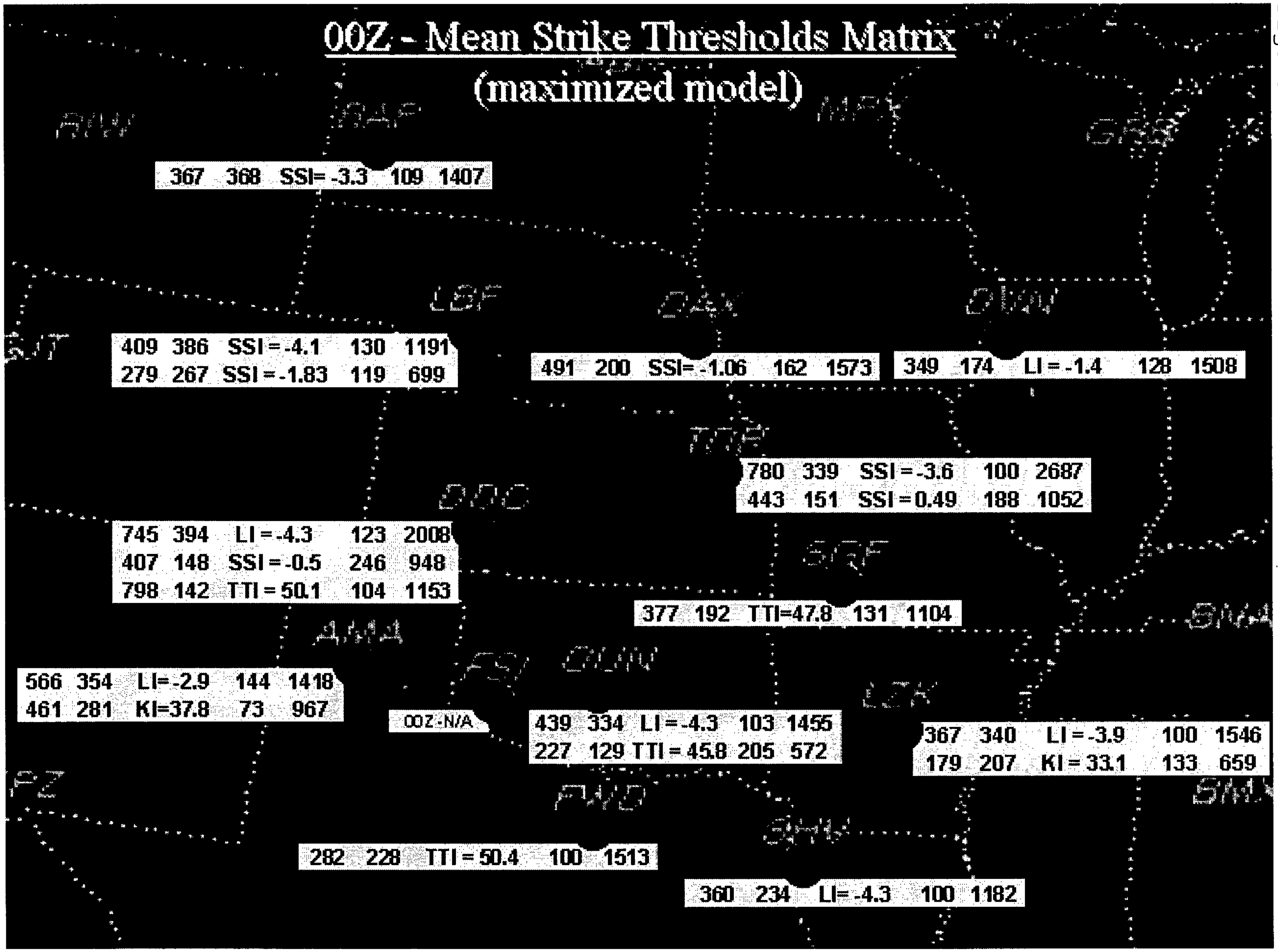
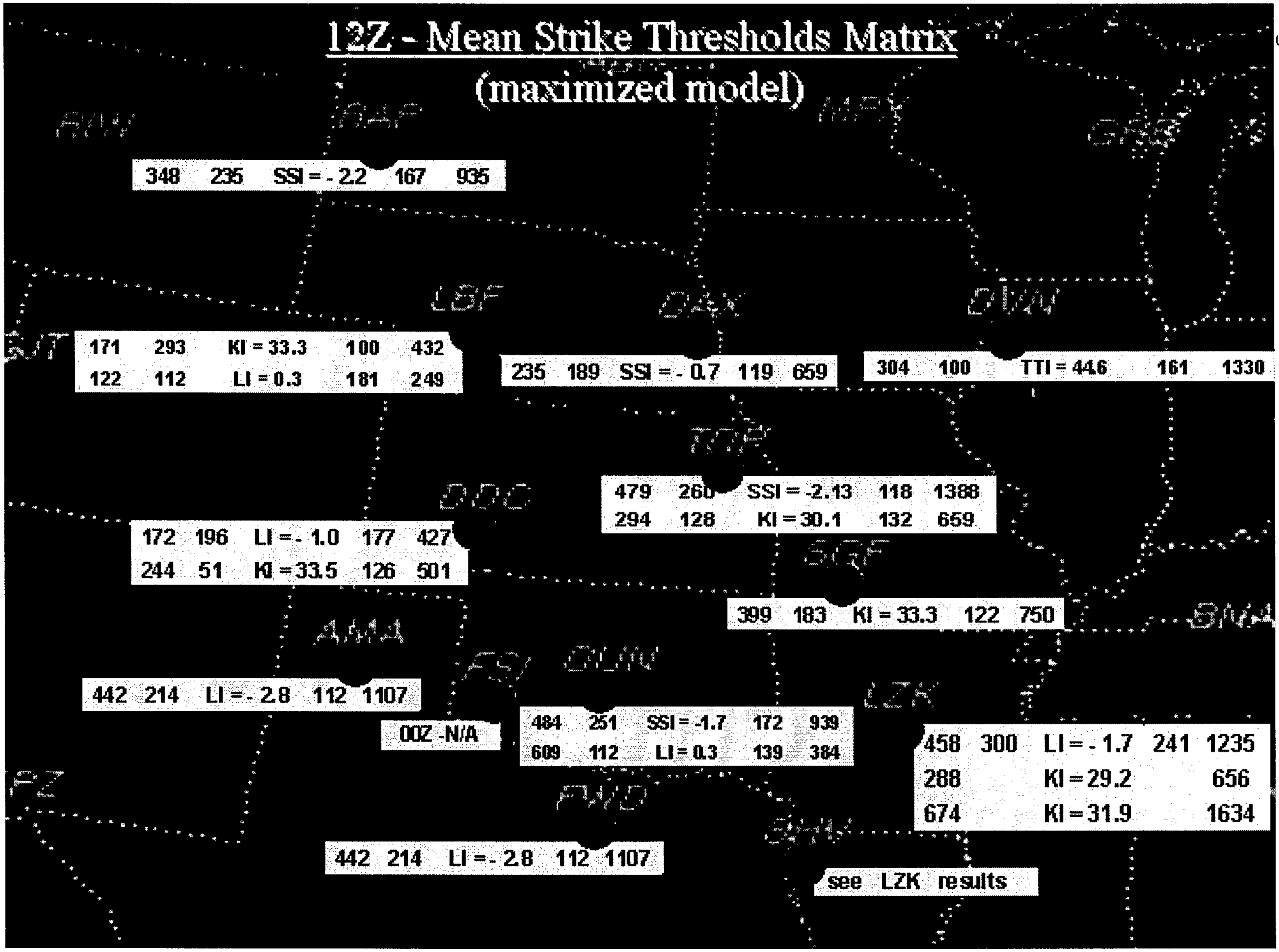

00Z - Mean Strike Thresholds Matrix
(maximized model)
RIW
RAP
MPX
GRB
367 368 SSI= -3.3 109 1407
LBF
OAX
DVN
409 386 SSI = -4.1 130 1191
279 267 SSI = -1.83 119 699
491 200 SSI= -1.06 162 1573
349 174 LI = -1.4 128 1508
TOP
780 339 SSI = -3.6 100 2687
443 151 SSI = 0.49 188 1052
DDC
SGF
745 394 LI = -4.3 123 2008
407 148 SSI = -0.5 246 948
798 142 TTI = 50.1 104 1153
377 192 TTI=47.8 131 1104
AMA
OUN
566 354 LI= -2.9 144 1418
461 281 KI=37.8 73 967
00Z -N/A
439 334 LI = -4.3 103 1455
227 129 TTI = 45.8 205 572
LZK
367 340 LI = -3.9 100 1546
179 207 KI = 33.1 133 659
FWD
SHV
282 228 TTI = 50.4 100 1513
360 234 LI= -4.3 100 1182

Figure A-8.   12Z mean strike thresholds results.



12Z - Mean Strike Thresholds Matrix
(maximized model)
RIW
RAP
MPX
GRB
LBF
OAX
DVN
DDC
TOP
SGF
AMA
FSI
OUN
LZK
FWD
SHV
BNA
348 235 SSI = - 2.2 167 935
171 293 KI = 33.3 100 432
122 112 LI = 0.3 181 249
235 189 SSI = - 0.7 119 659
304 100 TTI = 44.6 161 1330
479 260 SSI = -2.13 118 1388
294 128 KI = 30.1 132 659
172 196 LI = - 1.0 177 427
244 51 KI = 33.5 126 501
399 183 KI = 33.3 122 750
442 214 LI = - 2.8 112 1107
00Z -N/A
484 251 SSI = -1.7 172 939
609 112 LI = 0.3 139 384
458 300 LI = - 1.7 241 1235
288 KI = 29.2 656
674 KI = 31.9 1634
442 214 LI = - 2.8 112 1107
see LZK results

Table A-3. 00Z Lightning Probability Results.

	OUN (1009)		
	T-Storm	N	P
if	KI>25.2	605	0.56
&	LI<-1.1	402	0.63
&	KI>35	157	0.74
	No T-Storm	N	P
if	KI<25.2	404	0.75
&	TTI<46.7	293	0.8
&	KI<10.8	107	0.87

	LBF (1037)		
	T-Storm	N	P
if	SSI<1.1	585	0.69
&	KI>30.6	305	0.78
&	TTI>52	147	0.86
	No T-Storm	N	P
if	SSI>1.1	452	0.76
&	SSI>5.6	152	0.84

	OAX (800)		
	T-Storm	N	P
if	SSI<1.3	377	0.66
&	KI>30.5	209	0.78
&	TTI>50.1	101	0.87
	No T-Storm	N	P
if	SSI>1.3	423	0.73
&	LI>2.5	296	0.81
&	KI<11.1	106	0.91

	TOP (930)		
	T-Storm	N	P
if	SSI<2.2	550	0.63
&	KI>22.9	441	0.69
&	TTI>49	120	0.74
&	KI>35.2	130	0.82
	No T-Storm	N	P
if	SSI>2.2	380	0.76
&	KI<10.6	124	0.9

	RAP (946)		
	T-Storm	N	P
if	LI<1.1	459	0.75
&	SSI<-3.3	128	0.91
	No T-Storm	N	P
if	LI>1.1	487	0.73
&	LI>2.6	339	0.81
&	LI>6.7	111	0.88

	SGF (723)		
	T-Storm	N	P
if	KI>30.7	239	0.73
&	KI>34.2	121	0.84
	No T-Storm	N	P
if	KI<30.7	484	0.69
&	KI<13.35	158	0.88

Table A-3. 00Z Lightning Probability Results (cont.).

FWD (816)

	T-Storm	N	P
if	KI>30.5	346	0.61
&	KI>37.1	110	0.8

	No T-Storm	N	P
if	KI<30.5	470	0.75
&	TTI<41.2	132	0.92

DDC (1012)

	T-Storm	N	P
if	SSI< -0.5	546	0.69
&	KI>30.8	366	0.77
&	SSI< -2.4	248	0.82

	No T-Storm	N	P
if	SSI> -0.5	466	0.7
&	KI<25.0	261	0.78
&	SSI>4.3	132	0.86

DVN (731)

	T-Storm	N	P
if	KI>25.5	291	0.7
&	SSI< -0.3	133	0.86

	No T-Storm	N	P
if	KI<25.5	440	0.78
&	KI<17	310	0.83
&	LI>3.8	203	0.89
&	KI<3.6	103	0.9

LZK (1006)

	T-Storm	N	P
if	KI>27.3	491	0.65
&	TTI>45.7	255	0.8
&	KI>33.4	153	0.84

	No T-Storm	N	P
if	KI<27.3	515	0.77
&	LI>0.4	310	0.86
&	KI<11.9	157	0.89

SHV (749)

	T-Storm	N	P
if	TTI>44.1	451	0.6
&	KI>26.6	405	0.7
or	KI>35.9	116	0.84

	No T-Storm	N	P
if	TTI<44.1	298	0.79
&	LI>1.2	143	0.87

AMA (979)

	T-Storm	N	P
if	LI< -0.4	479	0.75
&	KI>38.4	139	0.87
or if	KI<38.4	340	0.69
&	TTI>51.7	154	0.77

	No T-Storm	N	P
if	LI> -0.4	500	0.72
&	LI>3.1	171	0.87
or if	LI<3.1	329	0.64
&	LI>1.4	109	0.71

Table A-3. 12Z Lightning Probability Results.

OUN (1003)

	T-Storm	N	P
if	KI>28.4	497	0.62
&	TTI>46.2	332	0.69
&	KI>35.4	167	0.79
	No T-Storm	N	P
if	KI<28.4	506	0.77
&	KI<15.5	182	0.89

LBF (1031)

	T-Storm	N	P
if	LI<4.0	691	0.49
&	KI>27.5	412	0.58
&	KI>32.9	162	0.66
	No T-Storm	N	P
if	LI>4.0	340	0.85
&	KI<12.0	100	0.95

OAX (800)

	T-Storm	N	P
if	KI>24.2	380	0.6
&	KI>32.0	188	0.71
	No T-Storm	N	P
if	KI<24.2	422	0.81
&	SSI>3.7	307	0.87
&	LI>10	100	0.93

TOP (930)

	T-Storm	N	P
if	KI>23.3	514	0.61
&	KI>32.9	239	0.77
&	TTI>47.9	137	0.84
	No T-Storm	N	P
if	KI<23.3	411	0.84
&	KI<7.7	136	0.94

RAP (927)

	T-Storm	N	P
if	LI<3.0	484	0.71
&	KI>25.4	292	0.78
&	LI< -0.6	119	0.88
	No T-Storm	N	P
if	LI>3.0	443	0.76
&	LI>8.3	136	0.93

SGF (714)

	T-Storm	N	P
if	KI>23.8	374	0.66
&	KI>33.2	143	0.87
	No T-Storm	N	P
if	KI<23.8	374	0.82
&	LI>7.2	120	0.97

Table A-3. 12Z Lightning Probability Results (cont.).

FWD (814)

	T-Storm	N	P
if	KI>27.8	440	0.58
&	KI>34.8	179	0.73
	No T-Storm	N	P
if	KI<27.8	374	0.82
&	SSI>2.95	174	0.93

DDC (994)

	T-Storm	N	P
if	KI>21.3	726	0.47
&	KI>32.7	306	0.61
&	KI>37.4	106	0.72
	No T-Storm	N	P
if	KI<21.3	268	0.89
&	TTI<40.5	118	0.93

DVN (728)

	T-Storm	N	P
if	KI>25.2	284	0.65
&	LI<1.1	169	0.76
	No T-Storm	N	P
if	KI<25.2	444	0.83
&	LI>3.2	327	0.89
&	TTI<38.2	218	0.94

LZK (1224)

	T-Storm	N	P
if	SSI<2.6	660	0.66
&	LI<1.4	543	0.72
&	KI>29.6	332	0.81
&	SSI<2.5	100	0.92
	No T-Storm	N	P
if	SSI>2.6	564	0.81
&	KI<10.9	273	0.94

SHV (748)

	T-Storm	N	P
if	KI>26.4	402	0.72
&	KI>33.5	179	0.87
	No T-Storm	N	P
if	KI>26.4	346	0.77
&	LI> -1.4	237	0.87
&	TTI<37	101	0.93

AMA (994)

	T-Storm	N	P
if	LI<1.9	646	0.57
&	KI>27.4	460	0.65
&	KI>35.1	156	0.76
	No T-Storm	N	P
if	LI>1.9	286	0.86
&	LI>5.1	162	0.94

Table A-3. 12Z Lightning Probability Results (cont.).

	FSI (195)		
	T-Storm	N	P
if	KI>30.6	86	0.62
	No T-Storm	N	P
if	KI>30.6	109	0.77
&	LI>2.2	56	0.89

Table A-4. 00Z Mean Lightning Strike Results.

mean	N	OUN (437)	N	mean
439	334	LI = -4.3	103	1455
227	129	TTI = 45.8	205	572

mean	N	LBF (516)	N	mean
409	386	SSI = -4.1	130	1191
279	267	SSI = -1.83	119	699

mean	N	OAX (362)	N	mean
491	200	SSI= -1.06	162	1573

mean	N	TOP (439)	N	mean
780	339	SSI = -3.6	100	2687
443	151	SSI = 0.49	188	1052

mean	N	RAP (477)	N	mean
367	368	SSI= -3.3	109	1407

mean	N	SGF (323)	N	mean
377	192	TTI=47.8	131	1104

mean	N	FWD (328)	N	mean
282	228	TTI = 50.4	100	1513

mean	N	DDC (517)	N	mean
745	394	LI = -4.3	123	2008
407	148	SSI = -0.5	246	948
798	142	TTI = 50.1	104	1153

mean	N	DVN (302)	N	mean
349	174	LI = -1.4	128	1508

mean	N	LZK (440)	N	mean
367	340	LI = -3.9	100	1546
179	207	KI = 33.1	133	659

mean	N	SHV (334)	N	mean
360	234	LI= -4.3	100	1182
169	110	TTI=45.5	124	530

mean	N	AMA (498)	N	mean
566	354	LI= -2.9	144	1418
461	281	KI=37.8	73	967

Table A-5. 12Z Mean Lightning Strike Results.

mean	N	OUN (423)	N	mean
484	251	SSI = -1.7	172	939
609	112	LI = 0.3	139	384

mean	N	LBF (393)	N	mean
171	293	KI = 33.3	100	432
122	112	LI = 0.3	181	249

mean	N	OAX (308)	N	mean
235	189	SSI = -0.68	119	659

mean	N	TOP (378)	N	mean
479	260	SSI = -2.13	118	1388
294	128	KI = 30.1	132	659

mean	N	RAP (402)	N	mean
348	235	SSI= -2.2	167	935

mean	N	SGF (305)	N	mean
399	183	KI=33.3	122	750

mean	N	FWD (326)	N	mean
442	214	LI = -2.8	112	1107

mean	N	DDC (373)	N	mean
172	196	LI = -1.0	177	427
244	51	KI = 33.5	126	501

mean	N	DVN (261)	N	mean
304	100	TTI = 44.6	161	1330

mean	N	LZK (541)	N	mean
458	300	LI = -1.7	241	1235
288		KI = 29.2		656
674		KI = 31.9		1634

mean	N	SHV ()	N	mean
		N/A See LZK Results		

mean	N	AMA ()	N	mean
		N/A See FWD Results		

Appendix B: Decision Tree Full Model Cross-Validation Results

The results of the full models are included for comparison reasons in appendix B to provide evidence of how well the decision tree results of this study cross-validate. Cross-validation means comparing a smaller study sample to the maximized database, and if similar results are found for the smaller study sample compared to the maximized sample, then the results cross-validate well and are considered significant. Full model in this case means that all of the indices were included in the decision tree model run which equated to a 30-40% smaller database due to missing data in a few of the stability indices that have already been determined as less significant predictors (KO, CAPE, and SWEAT). So the maximized model results do not include the less significant indices and therefore is "maximized" and 30-40% larger.

The following are some examples of very significant, almost identical, cross-validations from the textual summary output. Table B-1 compares the maximized model of 00Z LBF to the full model. Again the total number of cases included is in parenthesis next to the location identifier. There are 631/1037 = 0.61 or 39% fewer cases in the full model results for 00Z LBF, yet striking similarities exist between the first tree node threshold (SSI<1.1 for T-Storm or SSI>1.1 for No T-Storm), which is always deemed the most significant because the remaining indices all depend (inclusively) upon the condition that SSI<1.1 for T-Storm or SSI>1.1 for No T-Storm and therefore must first exist to be valid. Again, when the number of cases available drop to near 100 or less, accuracy becomes questionable. So KO<3.3 for the full model No T-Storm category only contains 51 cases and its inclusion in the maximized model results is not recommended. Next, compare Tables B-2 and B-3, and notice the similarities between the maximized tree model results versus the full model. The reader is encouraged to assess the cross-validations of the other locations as well as the 12Z sounding results of the full model in this appendix to the maximized model results in Appendix A.

Table B-1. Maximized versus Full classification tree results for 00Z LBF.

	LBF (1037)			VS.		LBF (631)		
	T-Storm	N	P			T-Storm	N	P
if	SSI<1.1	585	0.69		if	SSI<1.1	355	0.7
&	KI>30.6	305	0.78		&	SSI< -3.4	119	0.86
&	TTI>52	147	0.86					
	No T-Storm	N	P			No T-Storm	N	P
if	SSI>1.1	452	0.76		if	SSI>1.1	276	0.75
&	SSI>5.6	152	0.84		&	SSI>5.2	101	0.85
					&	KO<3.3	51	0.96

Table B-2. Maximized versus Full classification tree results for 00Z FWD.

	FWD (816)			VS.		FWD (471)		
	T-Storm	N	P			T-Storm	N	P
if	KI>30.5	346	0.61		if	KI>27.05	288	0.59
&	KI>37.1	110	0.8		&	TTI>41.10	223	0.68
&	TTI>47	108	0.65					
	No T-Storm	N	P			No T-Storm	N	P
if	KI<30.5	470	0.75		if	KI<27.05	183	0.78
&	TTI<41.2	132	0.92		&	TTI<41.10	61	0.92
&	KI<23.8	160	0.77					

Table B-3. Maximized versus Full classification tree results for 00Z LZK.

	LZK (1006)			VS.		LZK (723)		
	T-Storm	N	P			T-Storm	N	P
if	KI>27.3	491	0.65		if	KI>27.3	364	0.67
&	TTI>45.7	255	0.8		&	TTI>45.7	191	0.81
&	KI>33.4	153	0.84		&	TTI>49.6	60	0.95
	No T-Storm	N	P			No T-Storm	N	P
if	KI<27.3	515	0.77		if	KI<27.3	359	0.77
&	LI>0.4	310	0.86		&	LI>0.4	213	0.85
&	KI<11.9	157	0.89					

Again, typically when the available cases exist above 100 for both models, they cross-validate very well. Notice that the 00Z LZK results in Table B-3 are nearly identical, yet the full model for 00Z LZK is 28% smaller. The cross-validation results should be very compelling to skeptics and should be considered other than just coincidence.

Assessing how well the regression tree results cross-validate is a little different in that the regression tree model only includes the cases for CG lightning events only, thus sufficiently reducing the number of cases available for regression tree model fit. Also, the regression tree model results are not inclusive and not categorical, but instead independent and numerical. Similarly though is that the most important or significant index and threshold value is listed first because of the higher number of cases available. Table B-4 below is an example regression tree cross-validation for 00Z OUN with the significant similarities highlighted. There are approximately 25% fewer cases in the full model. Again, the results between the maximum and full models are strikingly similar.

Table B-4. Maximized versus Full regression tree results for 00Z OUN.

mean	N	OUN (437)	N	mean
439	334	LI = -4.3	103	1455
227	129	TTI = 45.8	205	572

mean	N	OUN (325)	N	mean
432	268	LI= -4.9	57	1722
162	102	TTI=45.8	166	598
424	126	KO= -10	40	1146

Table B-5. 00Z Full Model Classification Tree Results.

	OUN (705)		
	T-Storm	N	P
if	KI>27.8	370	0.61
&	SWEAT>230	175	0.72
	No T-Storm	N	P
if	KI<27.8	335	0.72
&	SWEAT<160	145	0.76
&	LI>3.0	106	0.82

	LBF (631)		
	T-Storm	N	P
if	SSI<1.1	355	0.7
&	SSI< -3.4	119	0.86
	No T-Storm	N	P
if	SSI>1.1	276	0.75
&	SSI>5.2	101	0.85
&	KO<3.3	51	0.96

	OAX (476)		
	T-Storm	N	P
if	SSI<3.1	278	0.64
&	TTI>46.9	107	0.65
&	KI>34.4	65	0.84
	No T-Storm	N	P
if	SSI>3.1	198	0.79
&	KI<15.3	100	0.89

	TOP (570)		
	T-Storm	N	P
if	SSI<0.5	277	0.69
&	KI>33.2	110	0.82
	No T-Storm	N	P
if	SSI>0.5	293	0.72
&	KI<25.2	203	0.81
&	KI>4.3	116	0.89

	RAP (499)		
	T-Storm	N	P
if	LI<1.4	251	0.74
&	LI< -0.8	137	0.87
&			
	No T-Storm	N	P
if	LI>1.4	248	0.79
&	SSI>4.3	141	0.86

SGF ()
N/A See LZK results

Table B-6. 00Z Full Model Classification Tree Results.

	FWD (471)		
	T-Storm	N	P
if	KI>27.05	288	0.59
&	TTI>41.10	223	0.68
	No T-Storm	N	P
if	KI<27.05	183	0.78
&	TTI<41.10	61	0.92

	DDC (569)		
	T-Storm	N	P
if	SSI< -1.0	286	0.75
&	KI>30.8	186	0.84
&	KI>35.0	128	0.88
	No T-Storm	N	P
if	SSI> -1.0	283	0.64
&	SWEAT<160	145	0.76
&	KO< -0.1	91	0.82

	DVN (437)		
	T-Storm	N	P
if	SSI<1.4	159	0.72
&	KI>30.1	98	0.83
	No T-Storm	N	P
if	SSI>1.4	278	0.75
&	KI<25.5	222	0.8
&	CAPE>5.2	146	0.87

	LZK (723)		
	T-Storm	N	P
if	KI>27.3	364	0.67
&	TTI>45.7	191	0.81
&	TTI>49.6	60	0.95
	No T-Storm	N	P
if	KI<27.3	359	0.77
&	LI>0.4	213	0.85

SHV ()
N/A See LZK results

	AMA (478)		
	T-Storm	N	P
if	LI< -0.4	255	0.77
&	SSI< -2.6	102	0.86
	No T-Storm	N	P
if	LI> -0.4	233	0.71
&	LI>2.1	104	0.83

Table B-7. 12Z Full Model Classification Tree Results.

OUN (723)

	T-Storm	N	P
if	KI>26.8	419	0.61
&	LI< -0.71	256	0.7
	No T-Storm	**N**	**P**
if	KI<26.8	304	0.79
&	CAPE<1812	253	0.88

LBF (521)

	T-Storm	N	P
if	LI<4.0	443	0.51
&	KI>25.4	316	0.58
&	LI> -10.5	266	0.61
&	CAPE>251.2	162	0.7
	No T-Storm	**N**	**P**
if	LI>4.0	178	0.85
&	KI<23.2	128	0.89

OAX (500)

	T-Storm	N	P
if	KI>24.2	257	0.63
&	KI>33.0	107	0.78
&	KI>36.5	51	0.84
	No T-Storm	**N**	**P**
if	KI<24.2	243	0.81
&	TTI<44.5	189	0.86

TOP (570)

	T-Storm	N	P
if	SSI<0.5	277	0.69
&	KI>33.2	110	0.82
	No T-Storm	**N**	**P**
if	SSI>0.5	293	0.72
&	KI<25.2	203	0.81
&	LI>4.3	116	0.89

RAP (477)

	T-Storm	N	P
if	LI<2.9	254	0.72
&	LI>26.9	144	0.82
&			
	No T-Storm	**N**	**P**
if	LI>2.9	223	0.78
&	KI<17.2	102	0.9

SGF (714)

N/A
See LZK
results

Table B-8. 12Z Full Model Classification Tree Results.

FWD (510)

	T-Storm	N	P
if	KI>27.8	291	0.58
&	KI>34.75	134	0.75
	No T-Storm	**N**	**P**
if	KI<27.8	219	0.81
&	KO> -13.3	168	0.9
&	TTI<42.55	105	0.96

DDC (676)

	T-Storm	N	P
if	KI>32.8	208	0.64
&	LI< -0.1	158	0.69
&	SSI< -3.7	56	0.71
	No T-Storm	**N**	**P**
if	KI<23	207	0.85
&	TTI<46.9	156	0.89

DVN (350)

	T-Storm	N	P
if	KI>23.2	180	0.66
&	CAPE>330	113	0.77
	No T-Storm	**N**	**P**
if	KI<23.2	170	0.84
&	TTI<38.2	90	0.94

LZK (719)

	T-Storm	N	P
if	SSI<1.2	371	0.73
&	KI>23	318	0.78
&	CAPE>679	255	0.82
&	KI>30.0	203	0.85
&	SSI< -2.6	69	0.93
	No T-Storm	**N**	**P**
if	SSI>1.2	348	0.76
&	KI>9.6	99	0.95

SHV (476)

	T-Storm	N	P
if	KI>33.5	134	0.87
&	KO< -11.0	84	0.93
	No T-Storm	**N**	**P**
if	KI<33.5	134	0.61
&	CAPE<701.5	108	0.86

AMA (591)

	T-Storm	N	P
if	LI<1.3	360	0.63
&	KI>28.6	255	0.71
&	KI>34.9	104	0.8
	No T-Storm	**N**	**P**
if	KI>1.3	231	0.79
&	KI<21.3	104	0.88

FSI (N/A)

T-Storm	N	P
NO SIGNIFICANT RESULTS		

Table B-9. 00Z Full Model Regression Tree Results.

mean	N	OUN (325)	N	mean
432	268	LI= -4.9	57	1722
162	102	TTI=45.8	166	598
424	126	KO= -10	40	1146

mean	N	LBF (253)	N	mean
186	200	KI=34.8	53	592
133	149	KO= -5.9	51	344

mean	N	OAX (219)	N	mean
526	176	SWEAT=298	43	1623
396	136	LI= -3.0	40	983
241	79	SWEAT=180	57	601

mean	N	TOP (272)	N	mean
914	230	SSI= -4.9	42	3172
552	136	KI=31.3	94	1438
270	72	CAPE=1100	64	869

mean	N	RAP (247)	N	mean
276	147	SSI= -1.2	100	1138

mean	N	SGF()	N	mean
		N/A		
		See LZK		
		results		

mean	N	FWD (328)	N	mean
316	186	SSI= -1.8	131	833
167	61	CAPE=1328	99	408

mean	N	DDC (317)	N	mean
656	238	SSI= -4.0	79	1913
492	196	KI=37.7	43	1423
377	151	SWEAT=260	45	878

mean	N	DVN (185)	N	mean
512	140	SSI= -1.5	45	2017
268	82	LI= -0.2	58	858

mean	N	LZK (327)	N	mean
261	313	SWEAT=252	66	1654
228	216	CAPE=2377	45	723

mean	N	SHV (261)	N	mean
415	149	SWEAT=238	112	1244

mean	N	AMA (261)	N	mean
550	208	SWEAT=312	53	1636
323	125	KI=35.4	83	893

Table B-10. 12Z Full Model Regression Tree Results.

mean	N	OUN (317)	N	mean
480	186	SSI= -1.8	131	833
519	60	CAPE=1987	71	1099
157	64	KI=33.4	70	519

mean	N	LBF (253)	N	mean
186	200	KI=34.8	53	592
133	149	KO= -5.9	51	344

mean	N	OAX (207)	N	mean
207	151	SWEAT=245.5	56	908
351	41	KI=32.4	110	1153

mean	N	TOP (261)	N	mean
378	175	SSI= -2.1	86	1343
207	114	CAPE=871	61	698

mean	N	RAP (233)	N	mean
142	115	SWEAT=164.5	118	475

mean	N	SGF ()	N	mean
		N/A		
		See LZK		
		results		

mean	N	FWD (211)	N	mean
493	159	LI= -3.7	52	1422
383	107	CAPE= 1676	52	720

mean	N	DDC (263)	N	mean
214	199	LI= -2.3	64	509

mean	N	DVN (146)	N	mean
541	102	CAPE=1541	44	1655
268	40	TTI=43.8	62	718

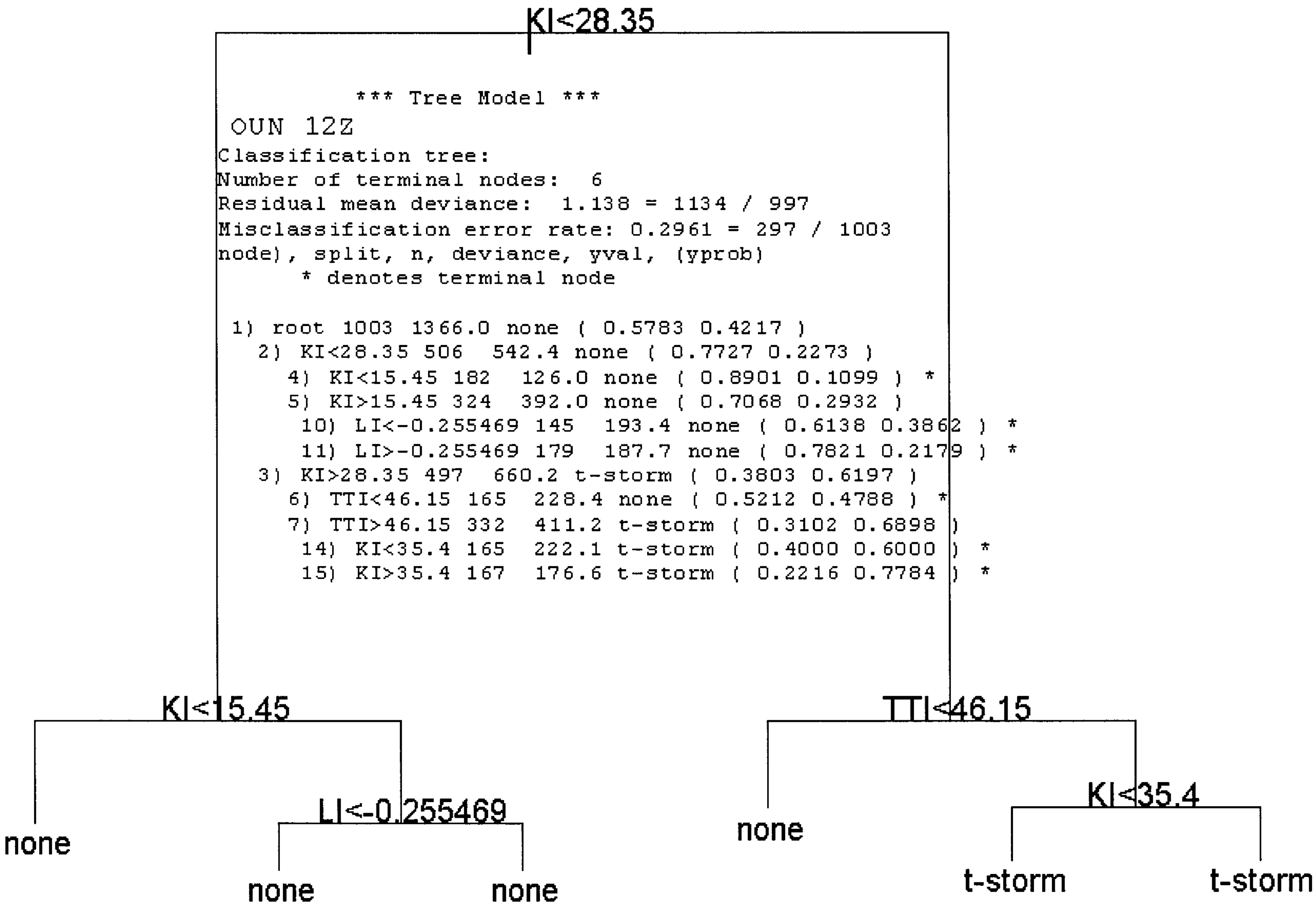
mean	N	LZK (354)	N	mean
372	148	LI= -1.1	206	1178
222	72	CAPE=576.4	76	514

mean	N	SHV (164)	N	mean
556	76	SSI= -1.9	76	1984

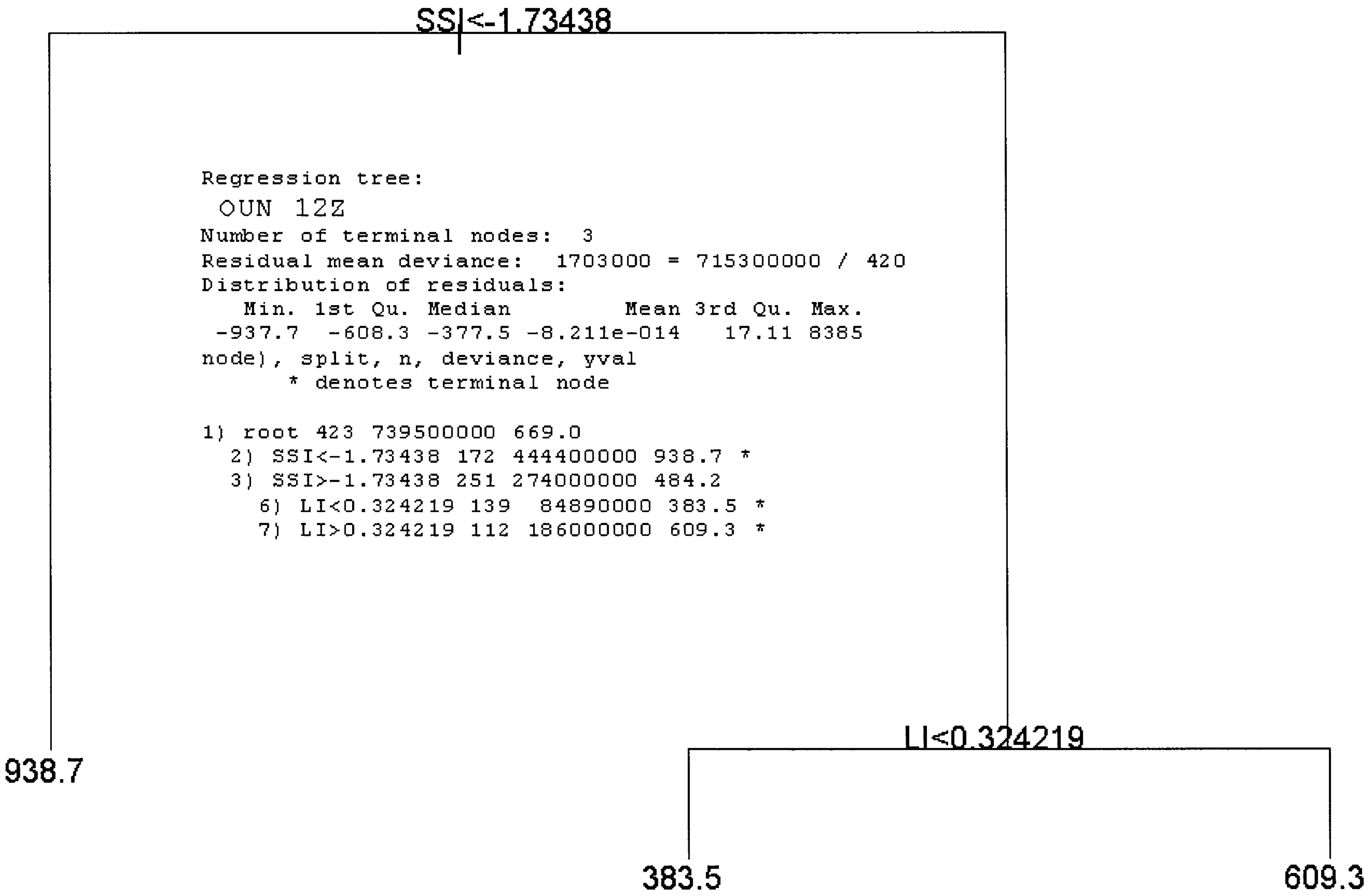
mean	N	AMA (274)	N	mean
204	126	SSI= -0.78	148	661

Appendix C: Final S-Plus® Decision Tree Output

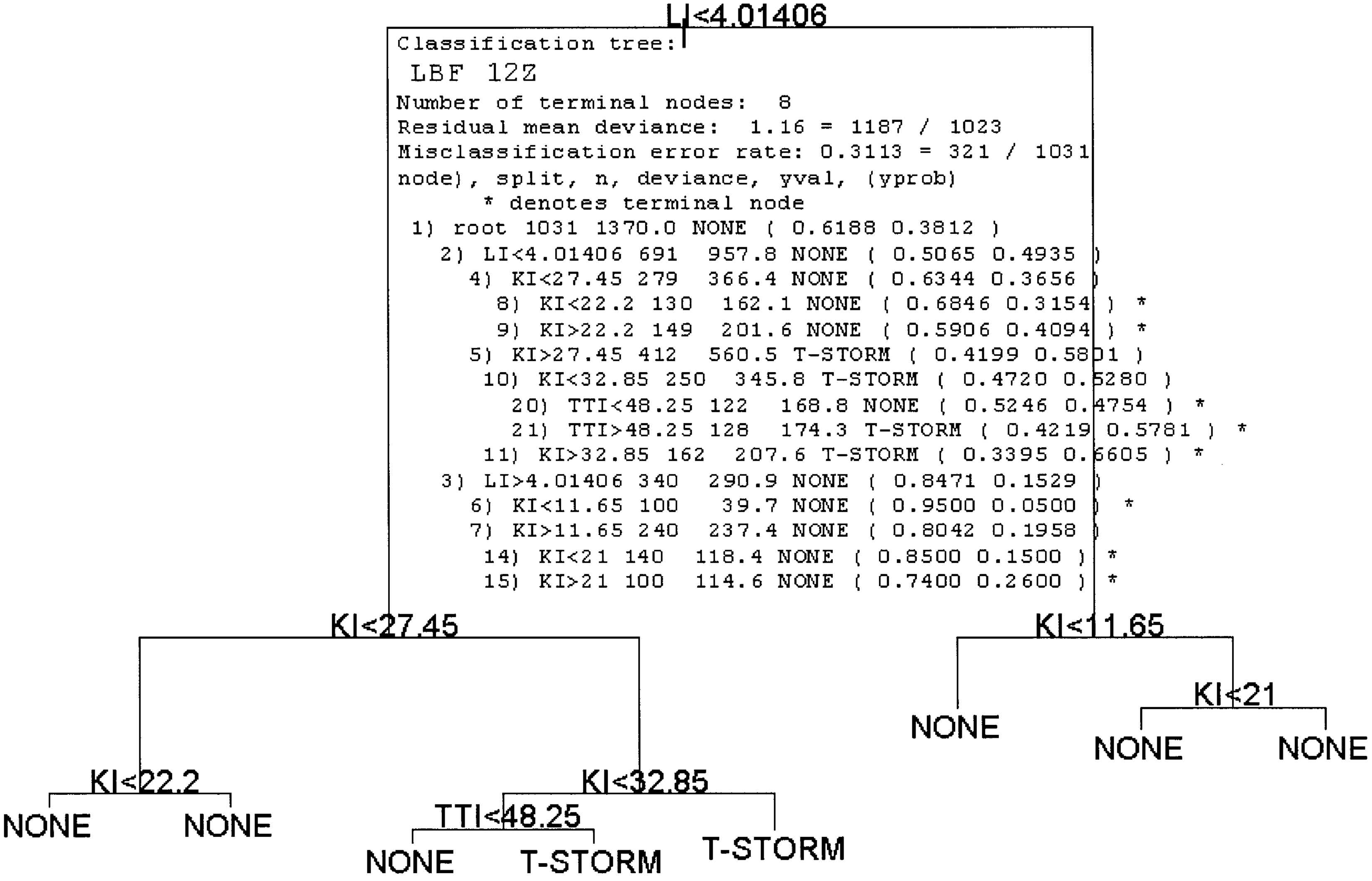
The graphics in the following pages are the combined, optimum, classification and regression tree outputs for each location in this study. The S-Plus graphical tree and textual summary output are combined and displayed. The information contained in each of these decision trees is the basis for which the results of the maximized model in Appendix A were developed.



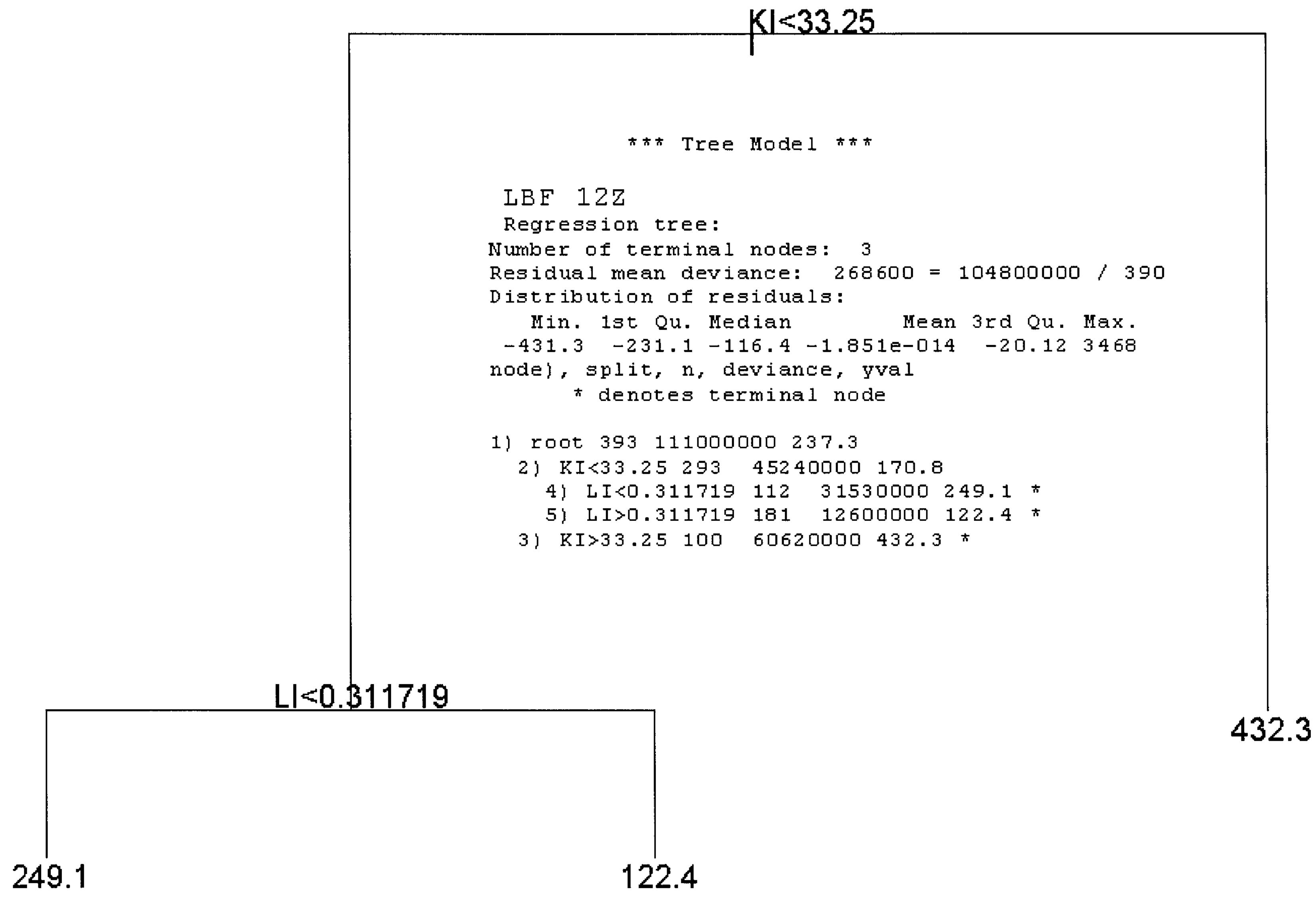
```
        *** Tree Model ***
 OUN 12Z
Classification tree:
Number of terminal nodes:  6
Residual mean deviance:  1.138 = 1134 / 997
Misclassification error rate: 0.2961 = 297 / 1003
node), split, n, deviance, yval, (yprob)
      * denotes terminal node

 1) root 1003 1366.0 none ( 0.5783 0.4217 )
   2) KI<28.35 506  542.4 none ( 0.7727 0.2273 )
     4) KI<15.45 182  126.0 none ( 0.8901 0.1099 ) *
     5) KI>15.45 324  392.0 none ( 0.7068 0.2932 )
      10) LI<-0.255469 145  193.4 none ( 0.6138 0.3862 ) *
      11) LI>-0.255469 179  187.7 none ( 0.7821 0.2179 ) *
   3) KI>28.35 497  660.2 t-storm ( 0.3803 0.6197 )
     6) TTI<46.15 165  228.4 none ( 0.5212 0.4788 ) *
     7) TTI>46.15 332  411.2 t-storm ( 0.3102 0.6898 )
      14) KI<35.4 165  222.1 t-storm ( 0.4000 0.6000 ) *
      15) KI>35.4 167  176.6 t-storm ( 0.2216 0.7784 ) *
```



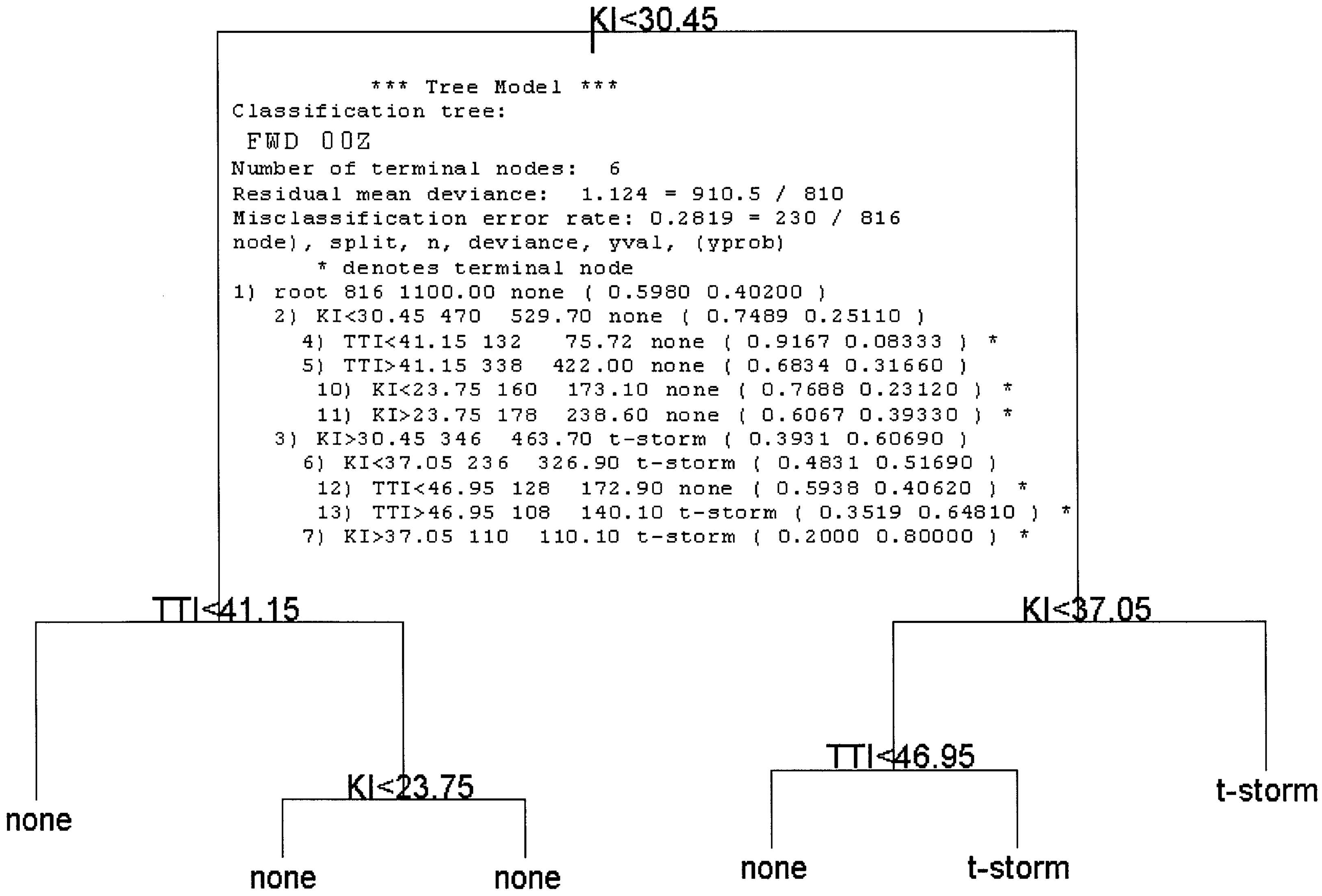
```
Regression tree:
 OUN 12Z
Number of terminal nodes:  3
Residual mean deviance:  1703000 = 715300000 / 420
Distribution of residuals:
   Min. 1st Qu. Median        Mean 3rd Qu. Max.
 -937.7  -608.3 -377.5 -8.211e-014   17.11 8385
node), split, n, deviance, yval
      * denotes terminal node

1) root 423 739500000 669.0
  2) SSI<-1.73438 172 444400000 938.7 *
  3) SSI>-1.73438 251 274000000 484.2
    6) LI<0.324219 139  84890000 383.5 *
    7) LI>0.324219 112 186000000 609.3 *
```



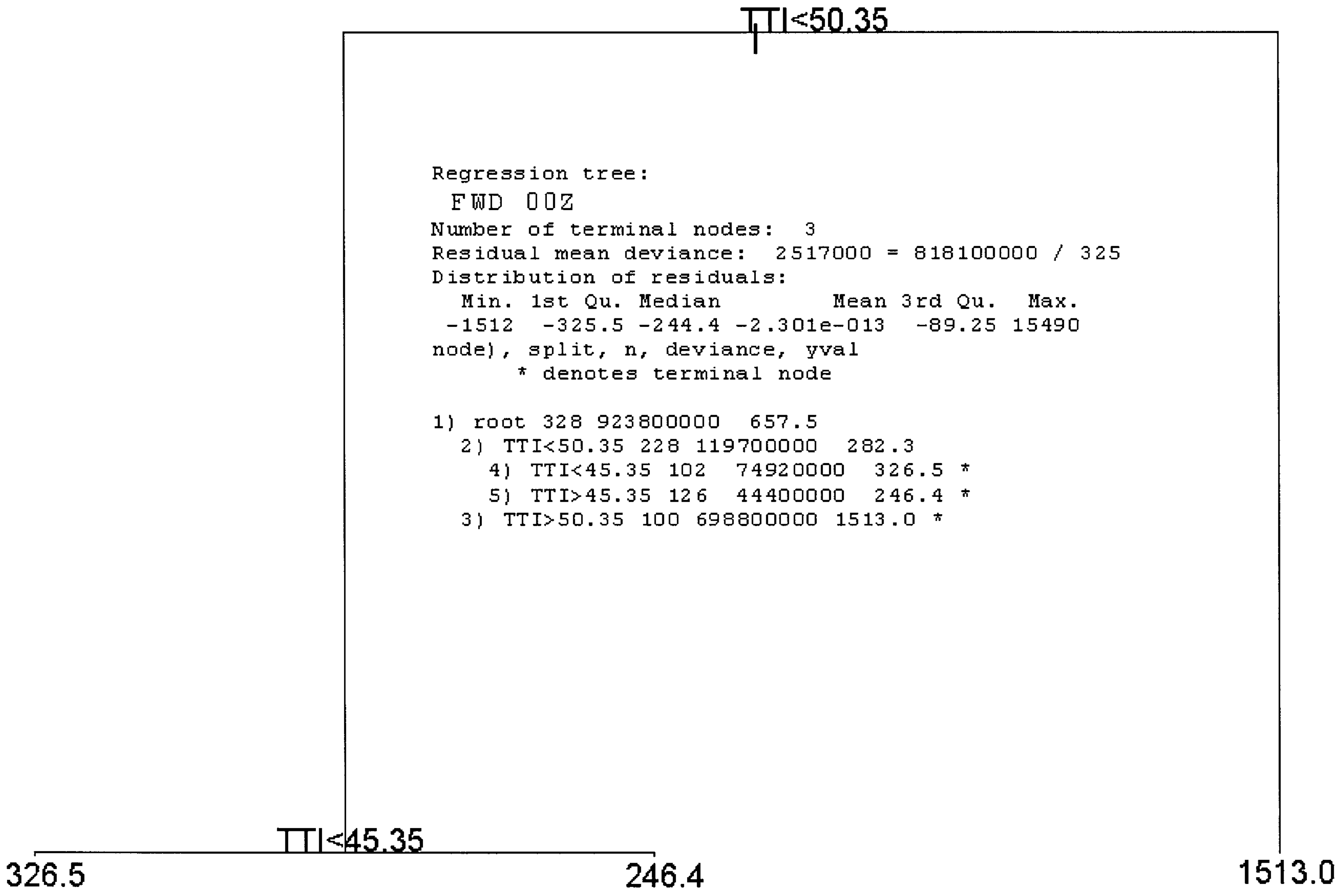
```
Classification tree:
 LBF 12Z
Number of terminal nodes:  8
Residual mean deviance:  1.16 = 1187 / 1023
Misclassification error rate: 0.3113 = 321 / 1031
node), split, n, deviance, yval, (yprob)
      * denotes terminal node
 1) root 1031 1370.0 NONE ( 0.6188 0.3812 )
   2) LI<4.01406 691  957.8 NONE ( 0.5065 0.4935 )
     4) KI<27.45 279  366.4 NONE ( 0.6344 0.3656 )
       8) KI<22.2 130  162.1 NONE ( 0.6846 0.3154 ) *
       9) KI>22.2 149  201.6 NONE ( 0.5906 0.4094 ) *
     5) KI>27.45 412  560.5 T-STORM ( 0.4199 0.5801 )
      10) KI<32.85 250  345.8 T-STORM ( 0.4720 0.5280 )
        20) TTI<48.25 122  168.8 NONE ( 0.5246 0.4754 ) *
        21) TTI>48.25 128  174.3 T-STORM ( 0.4219 0.5781 ) *
      11) KI>32.85 162  207.6 T-STORM ( 0.3395 0.6605 ) *
   3) LI>4.01406 340  290.9 NONE ( 0.8471 0.1529 )
     6) KI<11.65 100   39.7 NONE ( 0.9500 0.0500 ) *
     7) KI>11.65 240  237.4 NONE ( 0.8042 0.1958 )
      14) KI<21 140  118.4 NONE ( 0.8500 0.1500 ) *
      15) KI>21 100  114.6 NONE ( 0.7400 0.2600 ) *
```



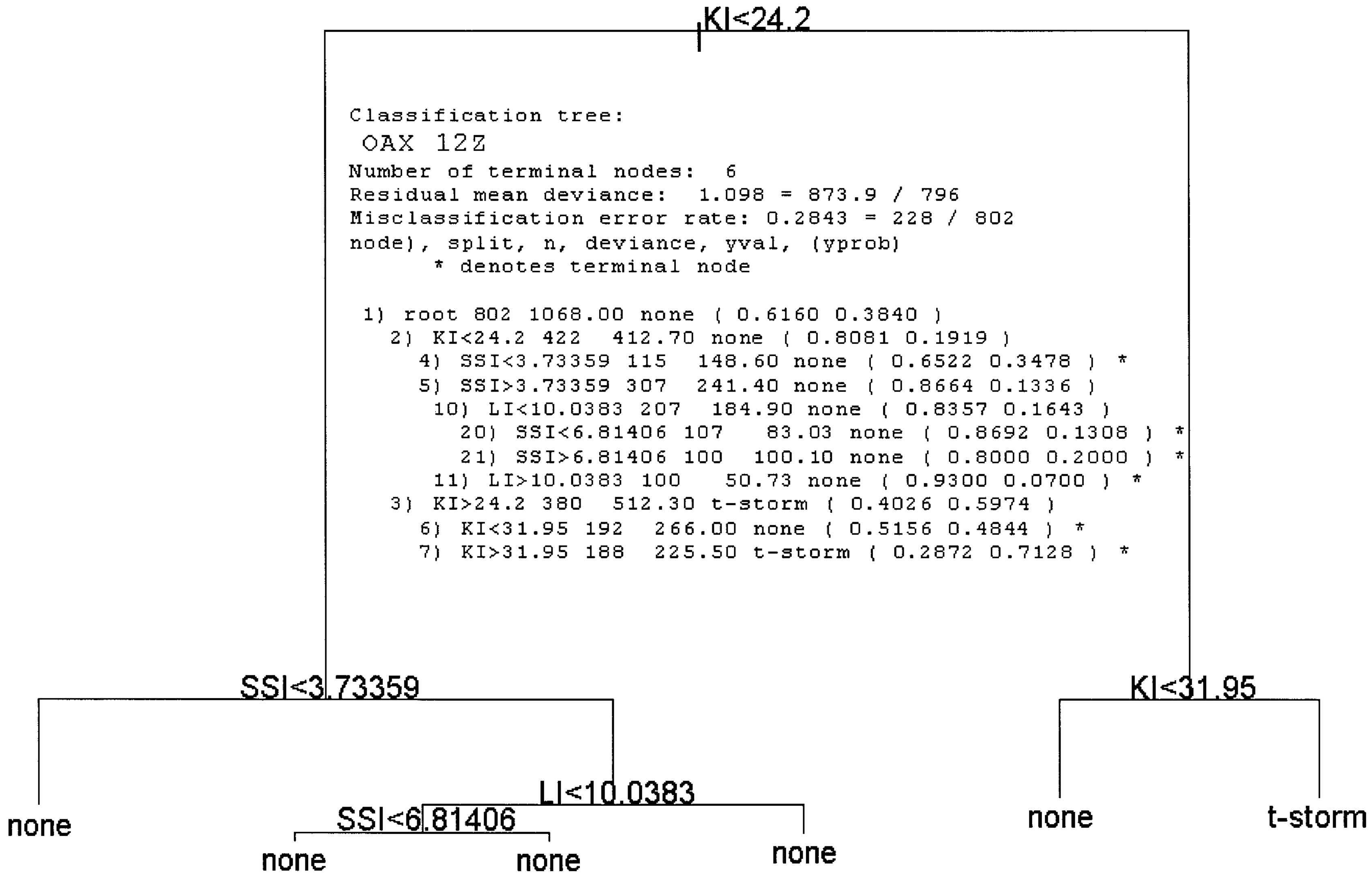
```
          *** Tree Model ***

 LBF 12Z
 Regression tree:
Number of terminal nodes:  3
Residual mean deviance:  268600 = 104800000 / 390
Distribution of residuals:
   Min. 1st Qu. Median        Mean 3rd Qu. Max.
 -431.3  -231.1 -116.4 -1.851e-014  -20.12 3468
node), split, n, deviance, yval
      * denotes terminal node

1) root 393 111000000 237.3
  2) KI<33.25 293  45240000 170.8
    4) LI<0.311719 112  31530000 249.1 *
    5) LI>0.311719 181  12600000 122.4 *
  3) KI>33.25 100  60620000 432.3 *
```



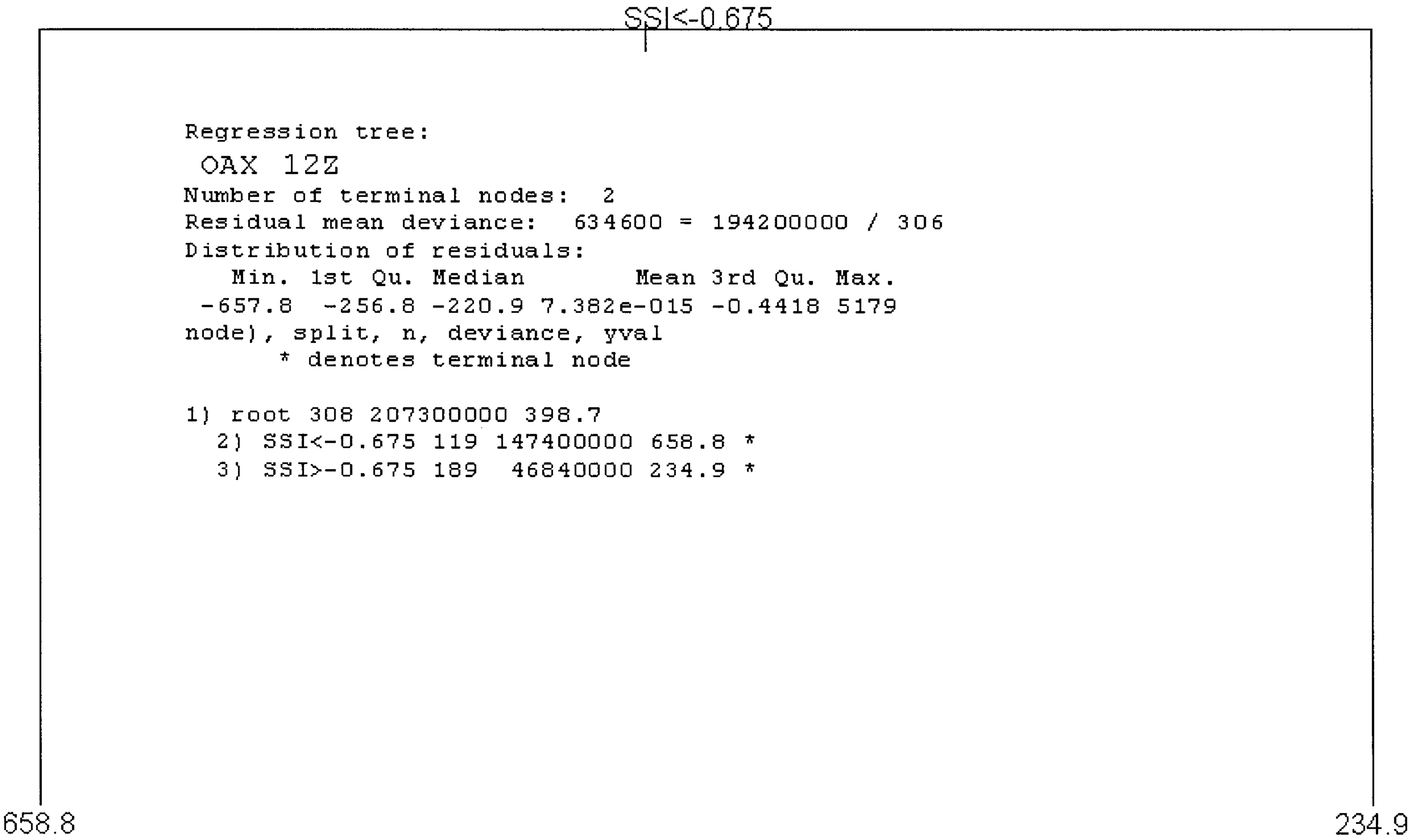
```
        *** Tree Model ***
Classification tree:
 FWD 00Z
Number of terminal nodes:  6
Residual mean deviance:  1.124 = 910.5 / 810
Misclassification error rate: 0.2819 = 230 / 816
node), split, n, deviance, yval, (yprob)
      * denotes terminal node
1) root 816 1100.00 none ( 0.5980 0.40200 )
   2) KI<30.45 470  529.70 none ( 0.7489 0.25110 )
     4) TTI<41.15 132   75.72 none ( 0.9167 0.08333 ) *
     5) TTI>41.15 338  422.00 none ( 0.6834 0.31660 )
      10) KI<23.75 160  173.10 none ( 0.7688 0.23120 ) *
      11) KI>23.75 178  238.60 none ( 0.6067 0.39330 ) *
   3) KI>30.45 346  463.70 t-storm ( 0.3931 0.60690 )
     6) KI<37.05 236  326.90 t-storm ( 0.4831 0.51690 )
      12) TTI<46.95 128  172.90 none ( 0.5938 0.40620 ) *
      13) TTI>46.95 108  140.10 t-storm ( 0.3519 0.64810 ) *
     7) KI>37.05 110  110.10 t-storm ( 0.2000 0.80000 ) *
```



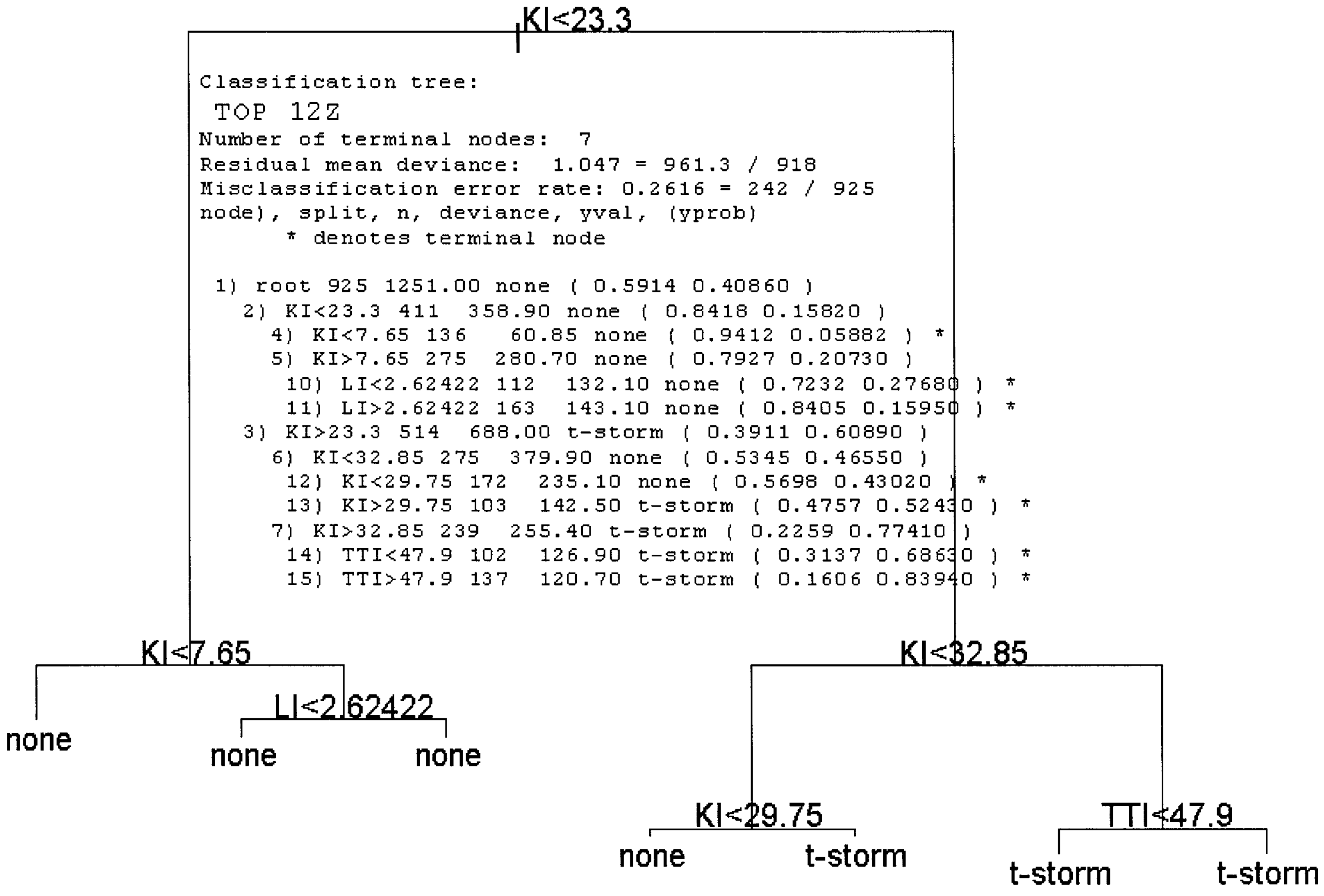
```
Regression tree:
 FWD 00Z
Number of terminal nodes:  3
Residual mean deviance:  2517000 = 818100000 / 325
Distribution of residuals:
  Min. 1st Qu. Median        Mean 3rd Qu.  Max.
 -1512  -325.5 -244.4 -2.301e-013  -89.25 15490
node), split, n, deviance, yval
      * denotes terminal node

1) root 328 923800000  657.5
  2) TTI<50.35 228 119700000  282.3
    4) TTI<45.35 102  74920000  326.5 *
    5) TTI>45.35 126  44400000  246.4 *
  3) TTI>50.35 100 698800000 1513.0 *
```



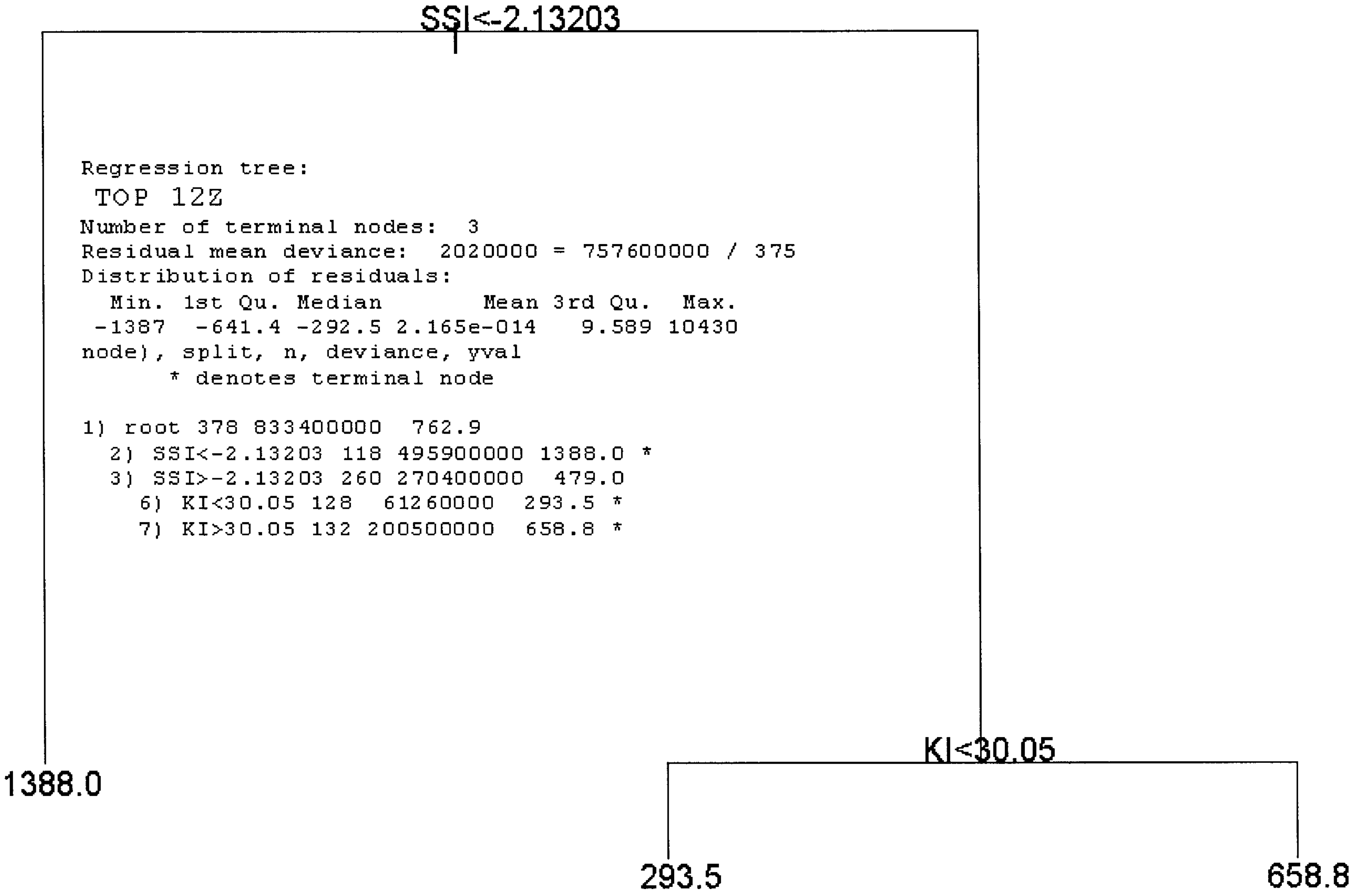
```
Classification tree:
 OAX 12Z
Number of terminal nodes:  6
Residual mean deviance:  1.098 = 873.9 / 796
Misclassification error rate: 0.2843 = 228 / 802
node), split, n, deviance, yval, (yprob)
      * denotes terminal node

 1) root 802 1068.00 none ( 0.6160 0.3840 )
   2) KI<24.2 422  412.70 none ( 0.8081 0.1919 )
     4) SSI<3.73359 115  148.60 none ( 0.6522 0.3478 ) *
     5) SSI>3.73359 307  241.40 none ( 0.8664 0.1336 )
      10) LI<10.0383 207  184.90 none ( 0.8357 0.1643 )
        20) SSI<6.81406 107   83.03 none ( 0.8692 0.1308 ) *
        21) SSI>6.81406 100  100.10 none ( 0.8000 0.2000 ) *
      11) LI>10.0383 100   50.73 none ( 0.9300 0.0700 ) *
   3) KI>24.2 380  512.30 t-storm ( 0.4026 0.5974 )
     6) KI<31.95 192  266.00 none ( 0.5156 0.4844 ) *
     7) KI>31.95 188  225.50 t-storm ( 0.2872 0.7128 ) *
```



```
Regression tree:
 OAX 12Z
Number of terminal nodes:  2
Residual mean deviance:  634600 = 194200000 / 306
Distribution of residuals:
   Min. 1st Qu. Median       Mean 3rd Qu. Max.
 -657.8  -256.8 -220.9 7.382e-015 -0.4418 5179
node), split, n, deviance, yval
      * denotes terminal node

1) root 308 207300000 398.7
  2) SSI<-0.675 119 147400000 658.8 *
  3) SSI>-0.675 189  46840000 234.9 *
```

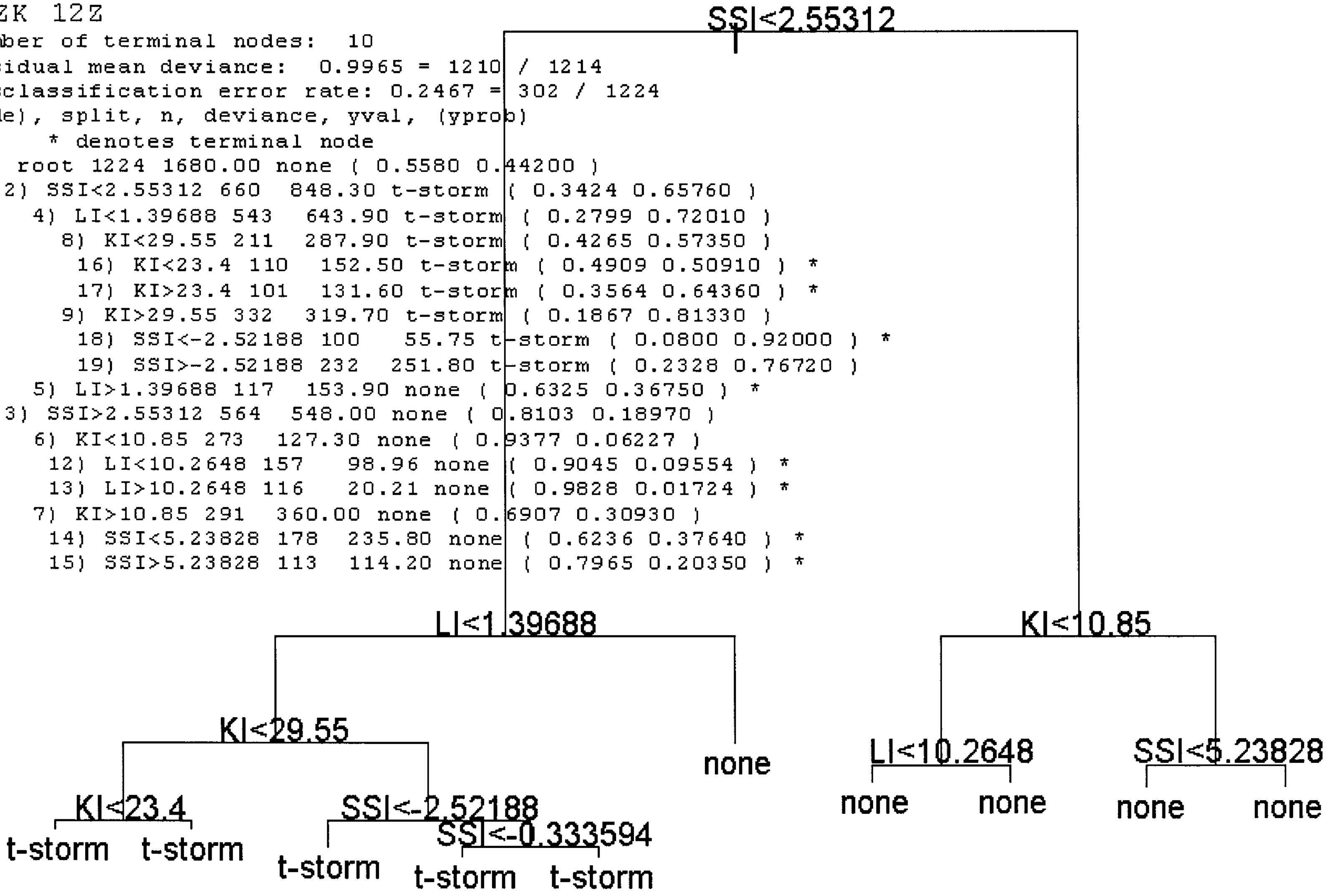


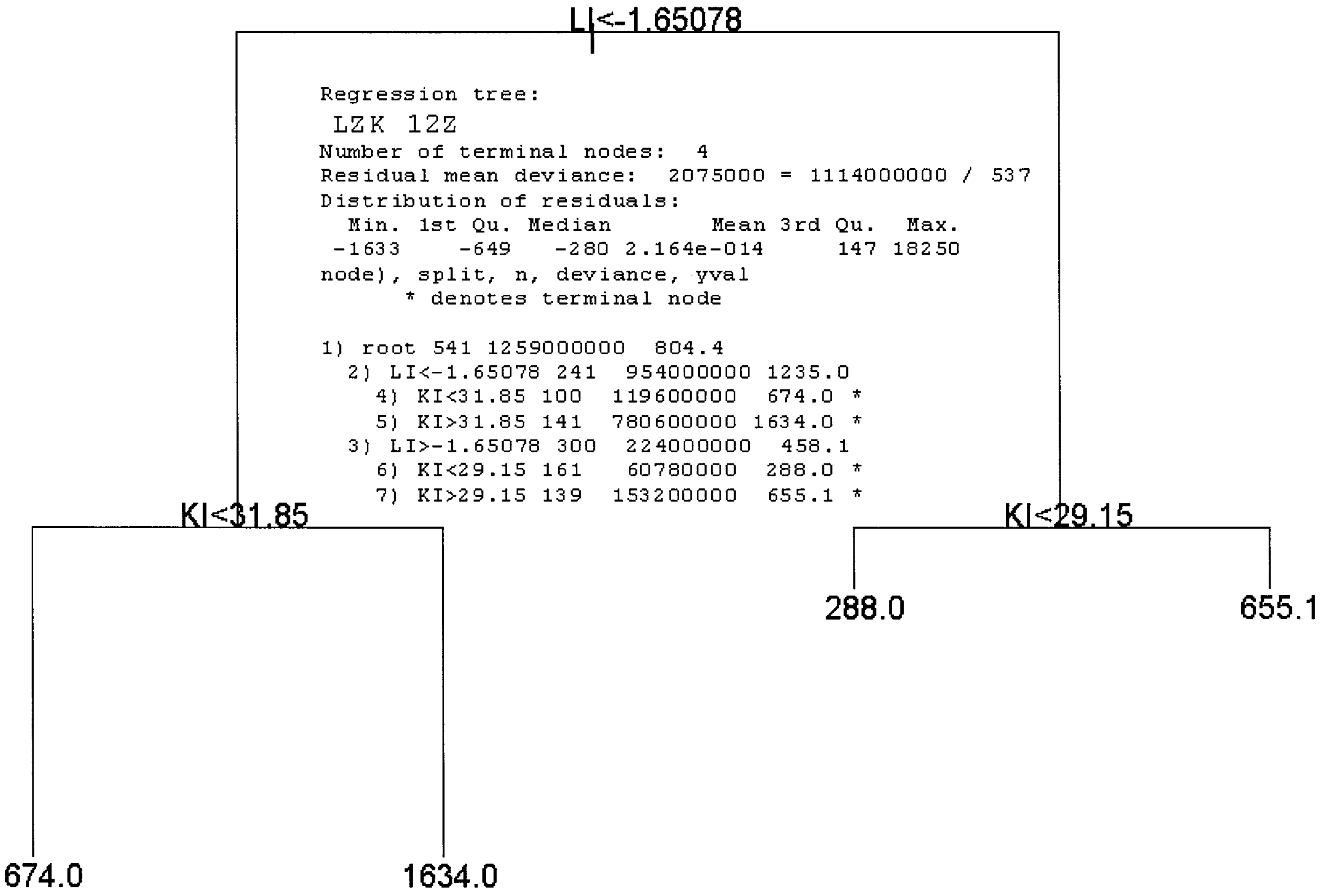
```
Classification tree:
 TOP 12Z
Number of terminal nodes:  7
Residual mean deviance:  1.047 = 961.3 / 918
Misclassification error rate: 0.2616 = 242 / 925
node), split, n, deviance, yval, (yprob)
      * denotes terminal node

 1) root 925 1251.00 none ( 0.5914 0.40860 )
   2) KI<23.3 411  358.90 none ( 0.8418 0.15820 )
     4) KI<7.65 136   60.85 none ( 0.9412 0.05882 ) *
     5) KI>7.65 275  280.70 none ( 0.7927 0.20730 )
      10) LI<2.62422 112  132.10 none ( 0.7232 0.27680 ) *
      11) LI>2.62422 163  143.10 none ( 0.8405 0.15950 ) *
   3) KI>23.3 514  688.00 t-storm ( 0.3911 0.60890 )
     6) KI<32.85 275  379.90 none ( 0.5345 0.46550 )
      12) KI<29.75 172  235.10 none ( 0.5698 0.43020 ) *
      13) KI>29.75 103  142.50 t-storm ( 0.4757 0.52430 ) *
     7) KI>32.85 239  255.40 t-storm ( 0.2259 0.77410 )
      14) TTI<47.9 102  126.90 t-storm ( 0.3137 0.68630 ) *
      15) TTI>47.9 137  120.70 t-storm ( 0.1606 0.83940 ) *
```



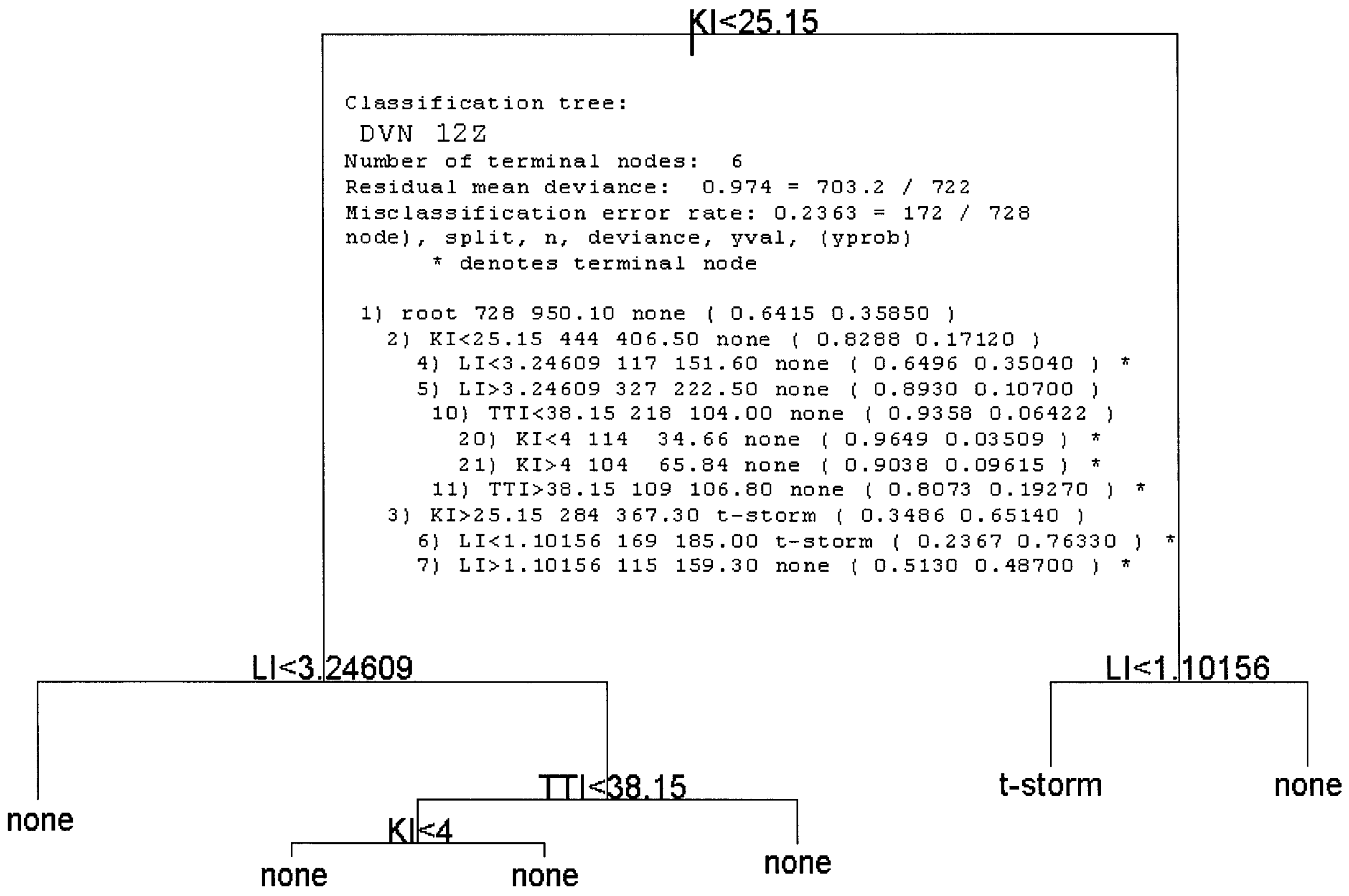
```
Regression tree:
 TOP 12Z
Number of terminal nodes:  3
Residual mean deviance:  2020000 = 757600000 / 375
Distribution of residuals:
  Min. 1st Qu. Median       Mean 3rd Qu.  Max.
 -1387  -641.4 -292.5 2.165e-014   9.589 10430
node), split, n, deviance, yval
      * denotes terminal node

1) root 378 833400000  762.9
  2) SSI<-2.13203 118 495900000 1388.0 *
  3) SSI>-2.13203 260 270400000  479.0
    6) KI<30.05 128  61260000  293.5 *
    7) KI>30.05 132 200500000  658.8 *
```

```
Classification tree:
 LZK 12Z
Number of terminal nodes:  10
Residual mean deviance:  0.9965 = 1210 / 1214
Misclassification error rate: 0.2467 = 302 / 1224
node), split, n, deviance, yval, (yprob)
      * denotes terminal node
 1) root 1224 1680.00 none ( 0.5580 0.44200 )
   2) SSI<2.55312 660  848.30 t-storm ( 0.3424 0.65760 )
     4) LI<1.39688 543  643.90 t-storm ( 0.2799 0.72010 )
       8) KI<29.55 211  287.90 t-storm ( 0.4265 0.57350 )
        16) KI<23.4 110  152.50 t-storm ( 0.4909 0.50910 ) *
        17) KI>23.4 101  131.60 t-storm ( 0.3564 0.64360 ) *
       9) KI>29.55 332  319.70 t-storm ( 0.1867 0.81330 )
        18) SSI<-2.52188 100   55.75 t-storm ( 0.0800 0.92000 ) *
        19) SSI>-2.52188 232  251.80 t-storm ( 0.2328 0.76720 )
     5) LI>1.39688 117  153.90 none ( 0.6325 0.36750 ) *
   3) SSI>2.55312 564  548.00 none ( 0.8103 0.18970 )
     6) KI<10.85 273  127.30 none ( 0.9377 0.06227 )
      12) LI<10.2648 157   98.96 none ( 0.9045 0.09554 ) *
      13) LI>10.2648 116   20.21 none ( 0.9828 0.01724 ) *
     7) KI>10.85 291  360.00 none ( 0.6907 0.30930 )
      14) SSI<5.23828 178  235.80 none ( 0.6236 0.37640 ) *
      15) SSI>5.23828 113  114.20 none ( 0.7965 0.20350 ) *
```

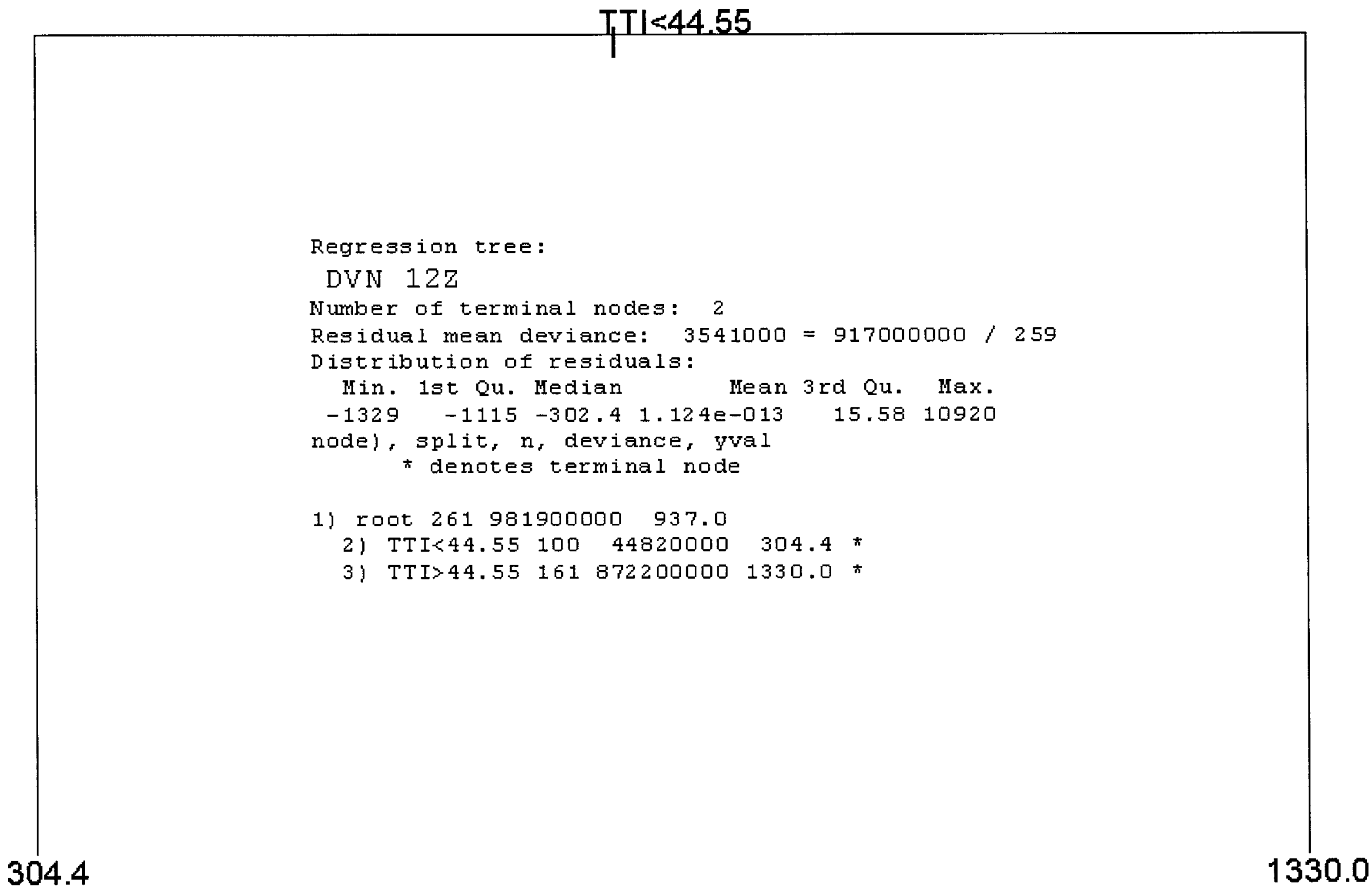




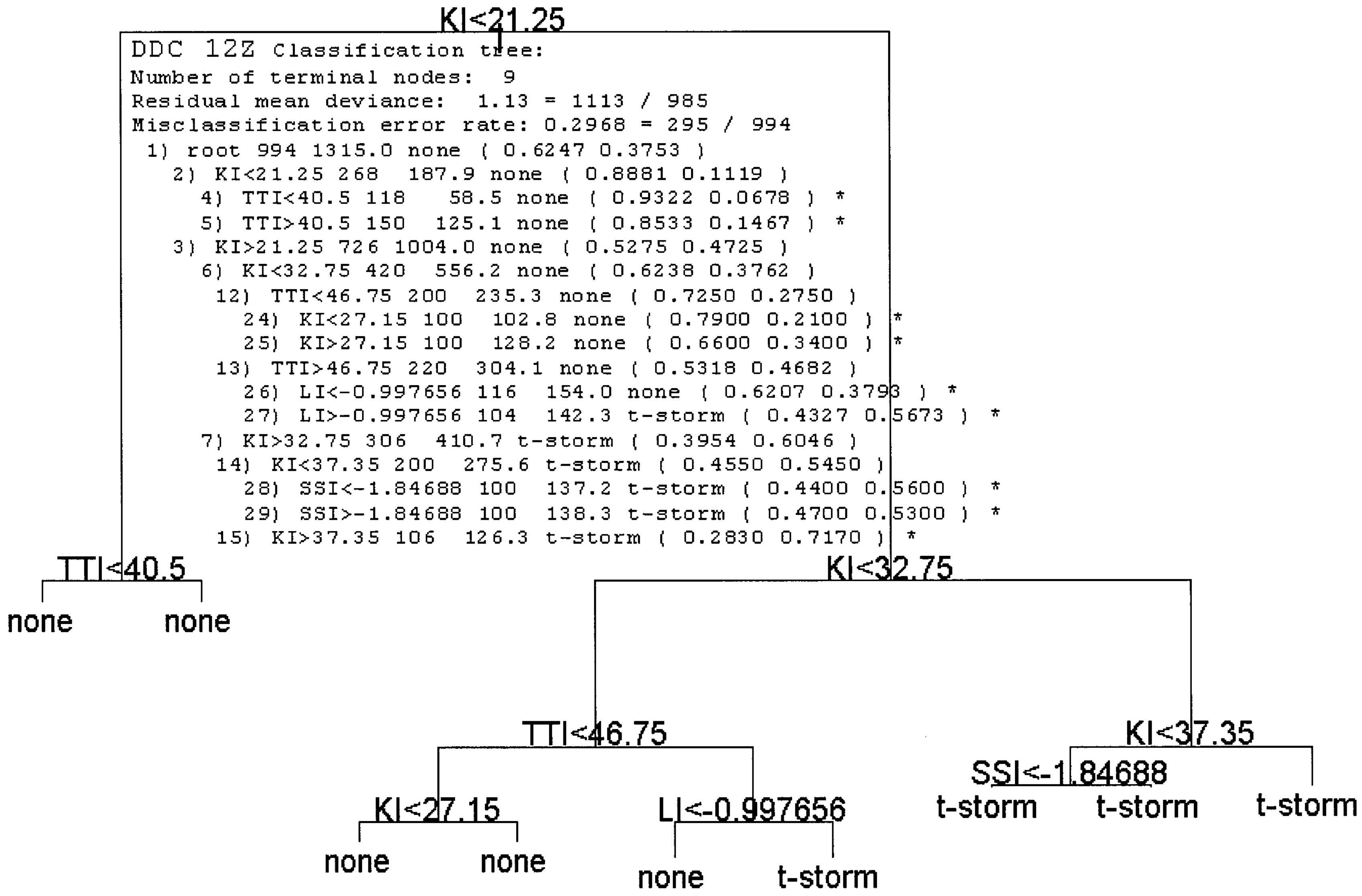
```
Regression tree:
 LZK 12Z
Number of terminal nodes:  4
Residual mean deviance:  2075000 = 1114000000 / 537
Distribution of residuals:
  Min. 1st Qu. Median       Mean 3rd Qu.  Max.
 -1633    -649   -280 2.164e-014     147 18250
node), split, n, deviance, yval
      * denotes terminal node

1) root 541 1259000000  804.4
  2) LI<-1.65078 241  954000000 1235.0
    4) KI<31.85 100  119600000  674.0 *
    5) KI>31.85 141  780600000 1634.0 *
  3) LI>-1.65078 300  224000000  458.1
    6) KI<29.15 161   60780000  288.0 *
    7) KI>29.15 139  153200000  655.1 *
```



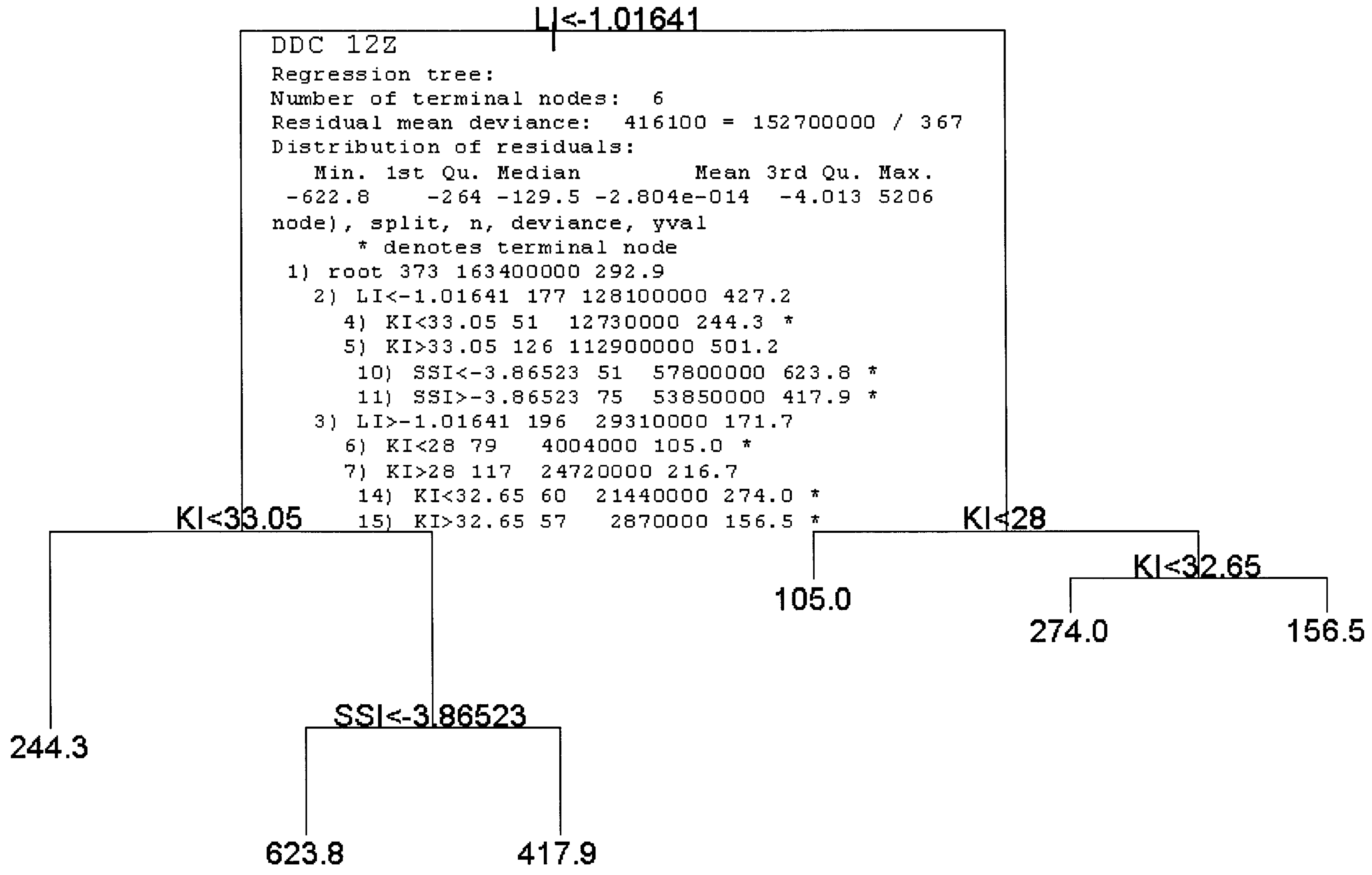
```
Classification tree:
 DVN 12Z
Number of terminal nodes:  6
Residual mean deviance:  0.974 = 703.2 / 722
Misclassification error rate: 0.2363 = 172 / 728
node), split, n, deviance, yval, (yprob)
      * denotes terminal node

 1) root 728 950.10 none ( 0.6415 0.35850 )
   2) KI<25.15 444 406.50 none ( 0.8288 0.17120 )
     4) LI<3.24609 117 151.60 none ( 0.6496 0.35040 ) *
     5) LI>3.24609 327 222.50 none ( 0.8930 0.10700 )
      10) TTI<38.15 218 104.00 none ( 0.9358 0.06422 )
        20) KI<4 114  34.66 none ( 0.9649 0.03509 ) *
        21) KI>4 104  65.84 none ( 0.9038 0.09615 ) *
      11) TTI>38.15 109 106.80 none ( 0.8073 0.19270 ) *
   3) KI>25.15 284 367.30 t-storm ( 0.3486 0.65140 )
     6) LI<1.10156 169 185.00 t-storm ( 0.2367 0.76330 ) *
     7) LI>1.10156 115 159.30 none ( 0.5130 0.48700 ) *
```



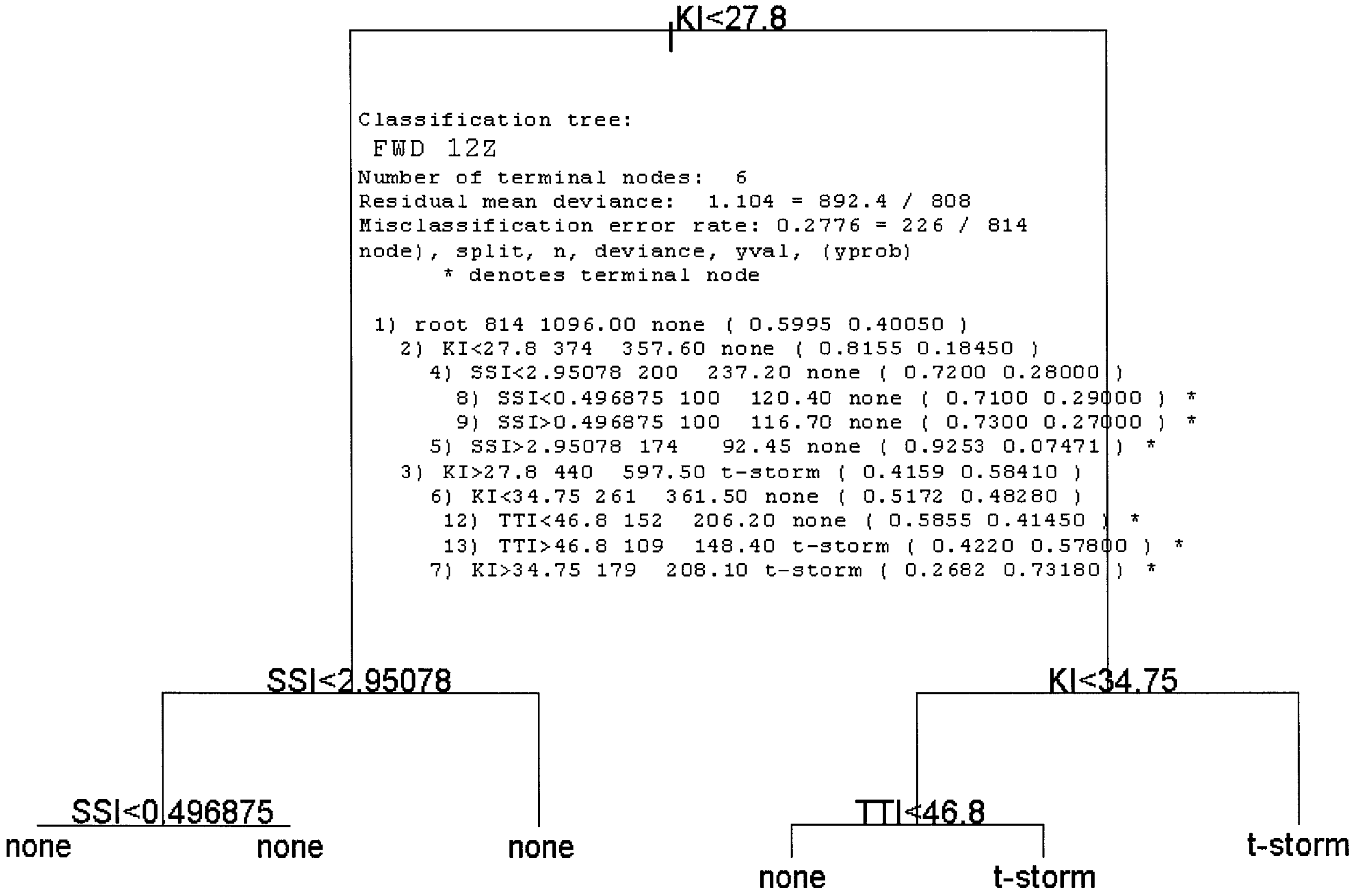
```
Regression tree:
 DVN 12Z
Number of terminal nodes:  2
Residual mean deviance:  3541000 = 917000000 / 259
Distribution of residuals:
  Min. 1st Qu. Median       Mean 3rd Qu.  Max.
 -1329   -1115 -302.4 1.124e-013   15.58 10920
node), split, n, deviance, yval
      * denotes terminal node

1) root 261 981900000  937.0
  2) TTI<44.55 100  44820000  304.4 *
  3) TTI>44.55 161 872200000 1330.0 *
```



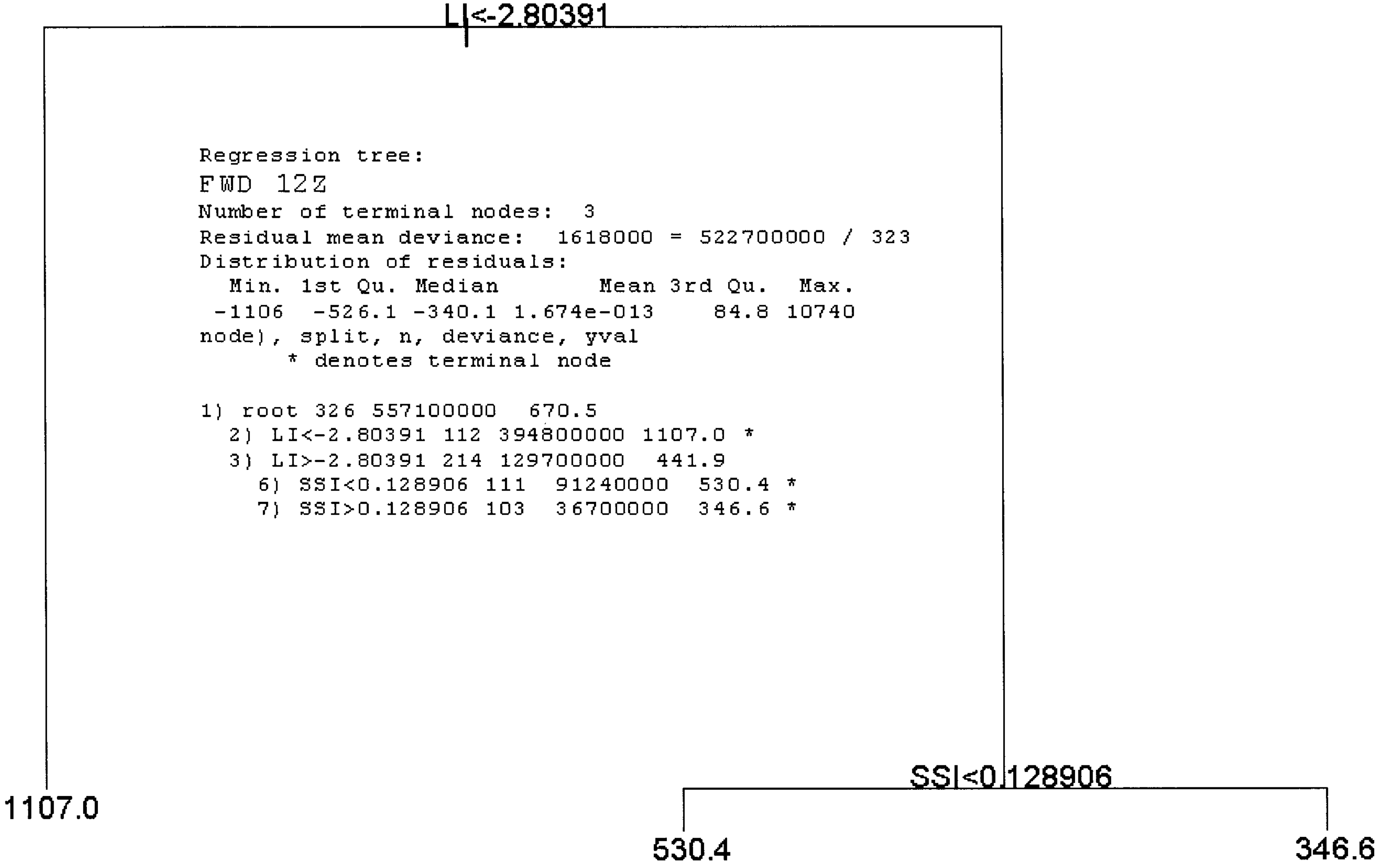
```
DDC 12Z Classification tree:
Number of terminal nodes:  9
Residual mean deviance:  1.13 = 1113 / 985
Misclassification error rate: 0.2968 = 295 / 994
 1) root 994 1315.0 none ( 0.6247 0.3753 )
   2) KI<21.25 268  187.9 none ( 0.8881 0.1119 )
     4) TTI<40.5 118   58.5 none ( 0.9322 0.0678 ) *
     5) TTI>40.5 150  125.1 none ( 0.8533 0.1467 ) *
   3) KI>21.25 726 1004.0 none ( 0.5275 0.4725 )
     6) KI<32.75 420  556.2 none ( 0.6238 0.3762 )
      12) TTI<46.75 200  235.3 none ( 0.7250 0.2750 )
        24) KI<27.15 100  102.8 none ( 0.7900 0.2100 ) *
        25) KI>27.15 100  128.2 none ( 0.6600 0.3400 ) *
      13) TTI>46.75 220  304.1 none ( 0.5318 0.4682 )
        26) LI<-0.997656 116  154.0 none ( 0.6207 0.3793 ) *
        27) LI>-0.997656 104  142.3 t-storm ( 0.4327 0.5673 ) *
     7) KI>32.75 306  410.7 t-storm ( 0.3954 0.6046 )
      14) KI<37.35 200  275.6 t-storm ( 0.4550 0.5450 )
        28) SSI<-1.84688 100  137.2 t-storm ( 0.4400 0.5600 ) *
        29) SSI>-1.84688 100  138.3 t-storm ( 0.4700 0.5300 ) *
      15) KI>37.35 106  126.3 t-storm ( 0.2830 0.7170 ) *
```



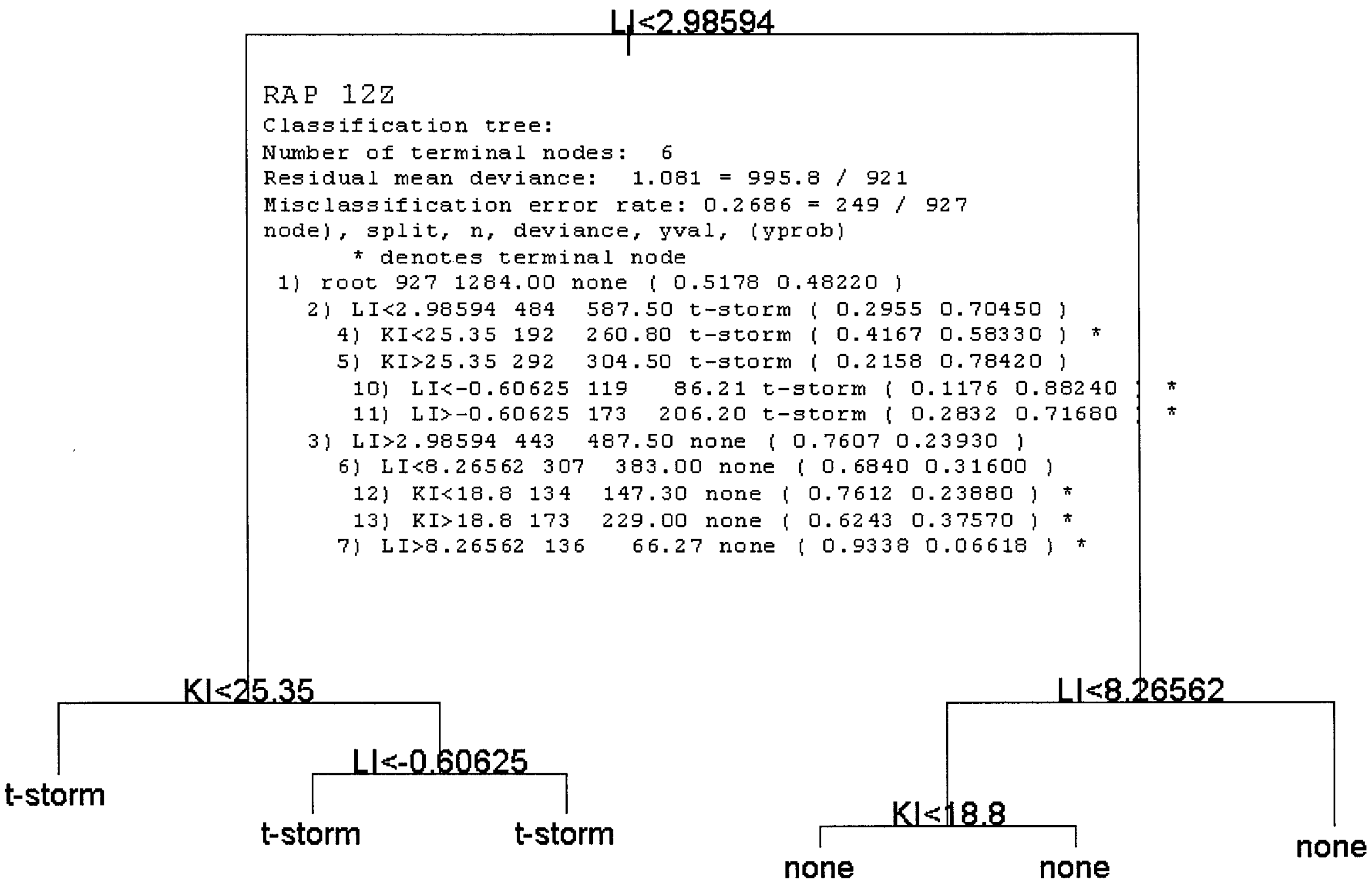
```
DDC 12Z
Regression tree:
Number of terminal nodes:  6
Residual mean deviance:  416100 = 152700000 / 367
Distribution of residuals:
   Min. 1st Qu. Median         Mean 3rd Qu. Max.
 -622.8    -264 -129.5 -2.804e-014  -4.013 5206
node), split, n, deviance, yval
      * denotes terminal node
 1) root 373 163400000 292.9
   2) LI<-1.01641 177 128100000 427.2
     4) KI<33.05 51  12730000 244.3 *
     5) KI>33.05 126 112900000 501.2
      10) SSI<-3.86523 51  57800000 623.8 *
      11) SSI>-3.86523 75  53850000 417.9 *
   3) LI>-1.01641 196  29310000 171.7
     6) KI<28 79   4004000 105.0 *
     7) KI>28 117  24720000 216.7
      14) KI<32.65 60  21440000 274.0 *
      15) KI>32.65 57   2870000 156.5 *
```



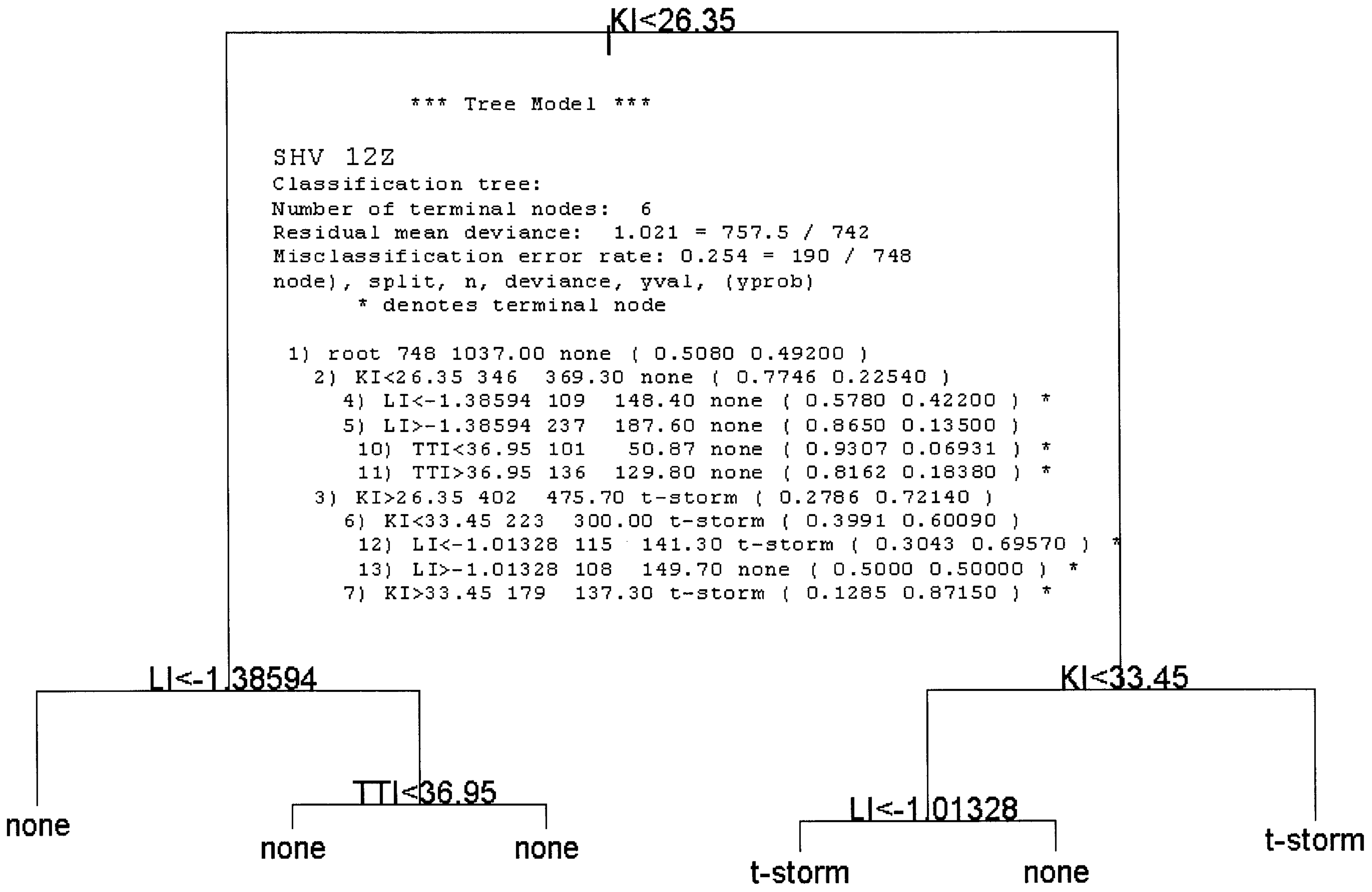
```
Classification tree:
 FWD 12Z
Number of terminal nodes:  6
Residual mean deviance:  1.104 = 892.4 / 808
Misclassification error rate: 0.2776 = 226 / 814
node), split, n, deviance, yval, (yprob)
      * denotes terminal node

 1) root 814 1096.00 none ( 0.5995 0.40050 )
   2) KI<27.8 374  357.60 none ( 0.8155 0.18450 )
     4) SSI<2.95078 200  237.20 none ( 0.7200 0.28000 )
       8) SSI<0.496875 100  120.40 none ( 0.7100 0.29000 ) *
       9) SSI>0.496875 100  116.70 none ( 0.7300 0.27000 ) *
     5) SSI>2.95078 174   92.45 none ( 0.9253 0.07471 ) *
   3) KI>27.8 440  597.50 t-storm ( 0.4159 0.58410 )
     6) KI<34.75 261  361.50 none ( 0.5172 0.48280 )
      12) TTI<46.8 152  206.20 none ( 0.5855 0.41450 ) *
      13) TTI>46.8 109  148.40 t-storm ( 0.4220 0.57800 ) *
     7) KI>34.75 179  208.10 t-storm ( 0.2682 0.73180 ) *
```



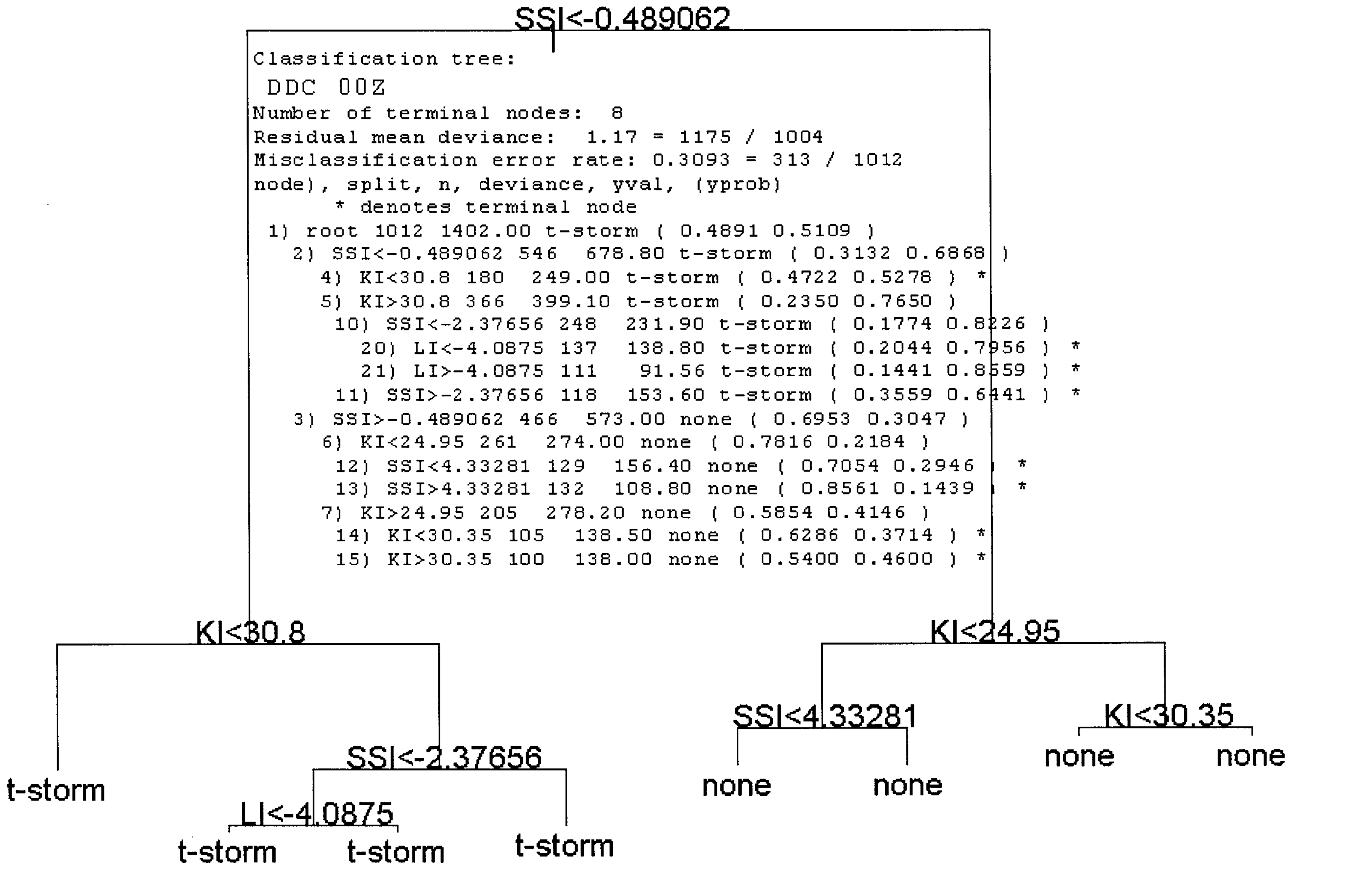
```
Regression tree:
FWD 12Z
Number of terminal nodes:  3
Residual mean deviance:  1618000 = 522700000 / 323
Distribution of residuals:
   Min. 1st Qu. Median       Mean 3rd Qu.  Max.
  -1106  -526.1 -340.1 1.674e-013    84.8 10740
node), split, n, deviance, yval
      * denotes terminal node

1) root 326 557100000  670.5
  2) LI<-2.80391 112 394800000 1107.0 *
  3) LI>-2.80391 214 129700000  441.9
    6) SSI<0.128906 111  91240000  530.4 *
    7) SSI>0.128906 103  36700000  346.6 *
```



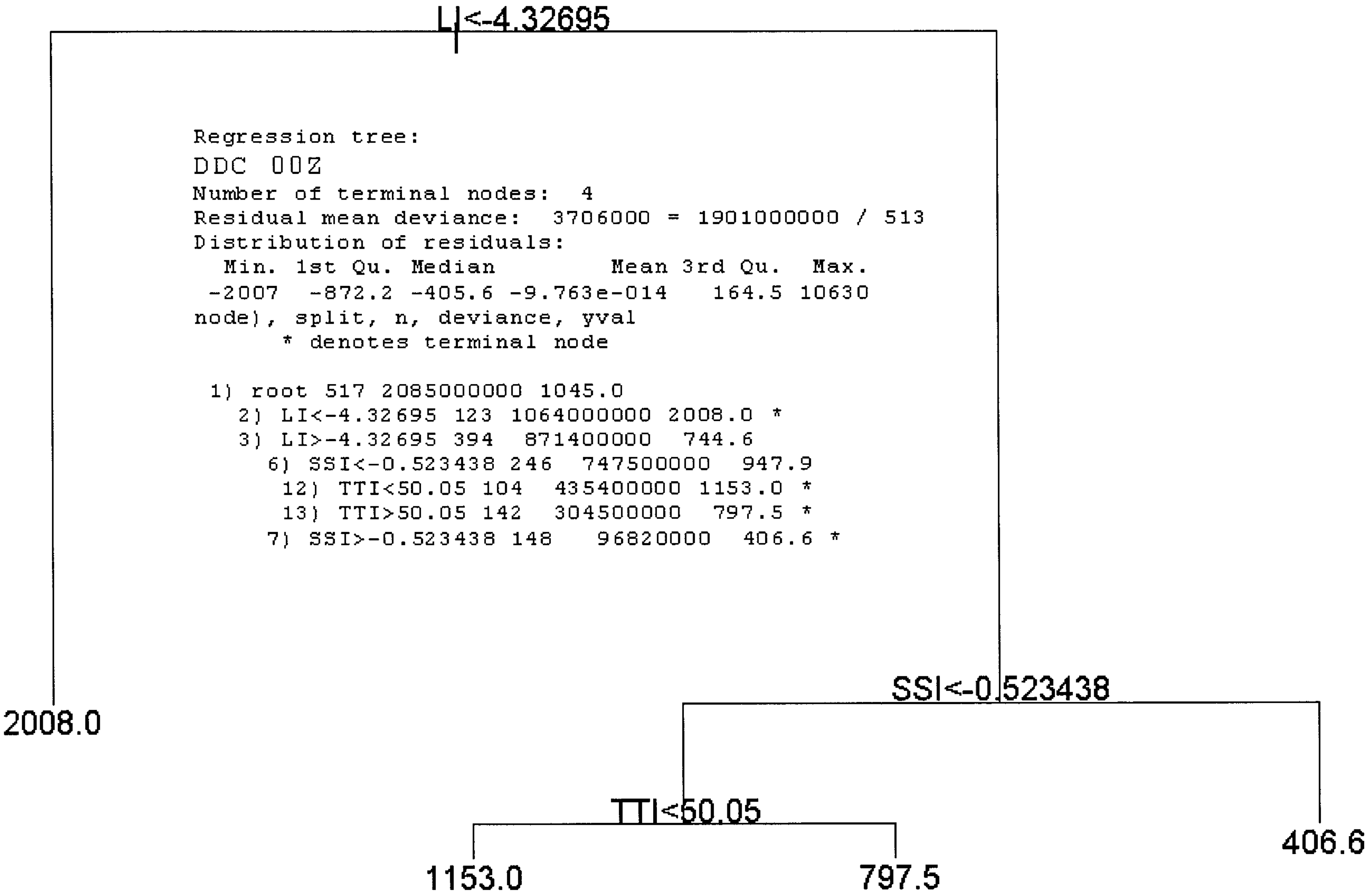
```
RAP 12Z
Classification tree:
Number of terminal nodes:  6
Residual mean deviance:  1.081 = 995.8 / 921
Misclassification error rate: 0.2686 = 249 / 927
node), split, n, deviance, yval, (yprob)
      * denotes terminal node
 1) root 927 1284.00 none ( 0.5178 0.48220 )
   2) LI<2.98594 484  587.50 t-storm ( 0.2955 0.70450 )
     4) KI<25.35 192  260.80 t-storm ( 0.4167 0.58330 ) *
     5) KI>25.35 292  304.50 t-storm ( 0.2158 0.78420 )
      10) LI<-0.60625 119   86.21 t-storm ( 0.1176 0.88240 ) *
      11) LI>-0.60625 173  206.20 t-storm ( 0.2832 0.71680 ) *
   3) LI>2.98594 443  487.50 none ( 0.7607 0.23930 )
     6) LI<8.26562 307  383.00 none ( 0.6840 0.31600 )
      12) KI<18.8 134  147.30 none ( 0.7612 0.23880 ) *
      13) KI>18.8 173  229.00 none ( 0.6243 0.37570 ) *
     7) LI>8.26562 136   66.27 none ( 0.9338 0.06618 ) *
```



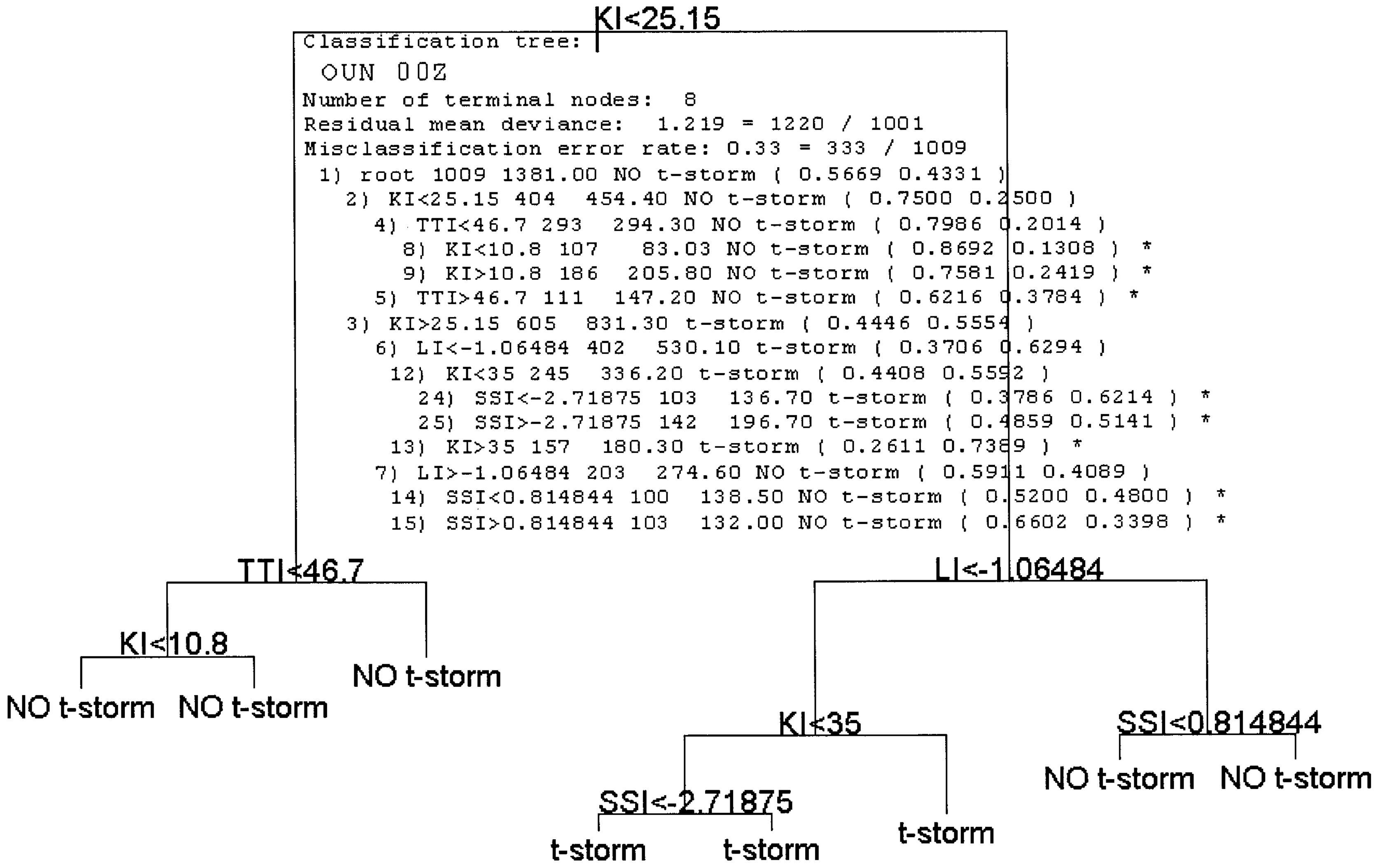
```
        *** Tree Model ***

SHV 12Z
Classification tree:
Number of terminal nodes:  6
Residual mean deviance:  1.021 = 757.5 / 742
Misclassification error rate: 0.254 = 190 / 748
node), split, n, deviance, yval, (yprob)
      * denotes terminal node

 1) root 748 1037.00 none ( 0.5080 0.49200 )
   2) KI<26.35 346  369.30 none ( 0.7746 0.22540 )
     4) LI<-1.38594 109  148.40 none ( 0.5780 0.42200 ) *
     5) LI>-1.38594 237  187.60 none ( 0.8650 0.13500 )
      10) TTI<36.95 101   50.87 none ( 0.9307 0.06931 ) *
      11) TTI>36.95 136  129.80 none ( 0.8162 0.18380 ) *
   3) KI>26.35 402  475.70 t-storm ( 0.2786 0.72140 )
     6) KI<33.45 223  300.00 t-storm ( 0.3991 0.60090 )
      12) LI<-1.01328 115  141.30 t-storm ( 0.3043 0.69570 ) *
      13) LI>-1.01328 108  149.70 none ( 0.5000 0.50000 ) *
     7) KI>33.45 179  137.30 t-storm ( 0.1285 0.87150 ) *
```



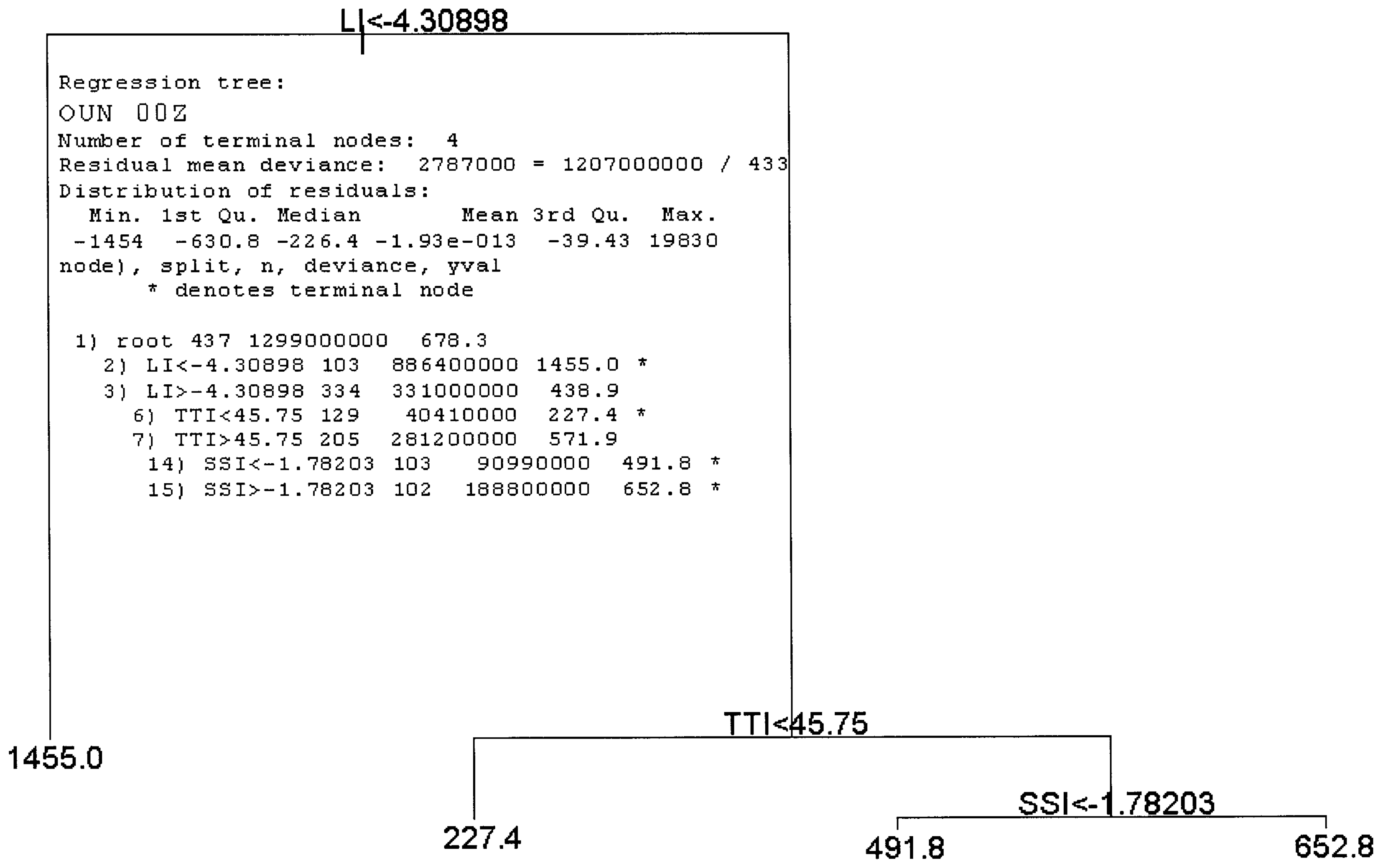
```
Classification tree:
 DDC 00Z
Number of terminal nodes:  8
Residual mean deviance:  1.17 = 1175 / 1004
Misclassification error rate: 0.3093 = 313 / 1012
node), split, n, deviance, yval, (yprob)
      * denotes terminal node
 1) root 1012 1402.00 t-storm ( 0.4891 0.5109 )
   2) SSI<-0.489062 546  678.80 t-storm ( 0.3132 0.6868 )
     4) KI<30.8 180  249.00 t-storm ( 0.4722 0.5278 ) *
     5) KI>30.8 366  399.10 t-storm ( 0.2350 0.7650 )
      10) SSI<-2.37656 248  231.90 t-storm ( 0.1774 0.8226 )
        20) LI<-4.0875 137  138.80 t-storm ( 0.2044 0.7956 ) *
        21) LI>-4.0875 111   91.56 t-storm ( 0.1441 0.8559 ) *
      11) SSI>-2.37656 118  153.60 t-storm ( 0.3559 0.6441 ) *
   3) SSI>-0.489062 466  573.00 none ( 0.6953 0.3047 )
     6) KI<24.95 261  274.00 none ( 0.7816 0.2184 )
      12) SSI<4.33281 129  156.40 none ( 0.7054 0.2946 ) *
      13) SSI>4.33281 132  108.80 none ( 0.8561 0.1439 ) *
     7) KI>24.95 205  278.20 none ( 0.5854 0.4146 )
      14) KI<30.35 105  138.50 none ( 0.6286 0.3714 ) *
      15) KI>30.35 100  138.00 none ( 0.5400 0.4600 ) *
```



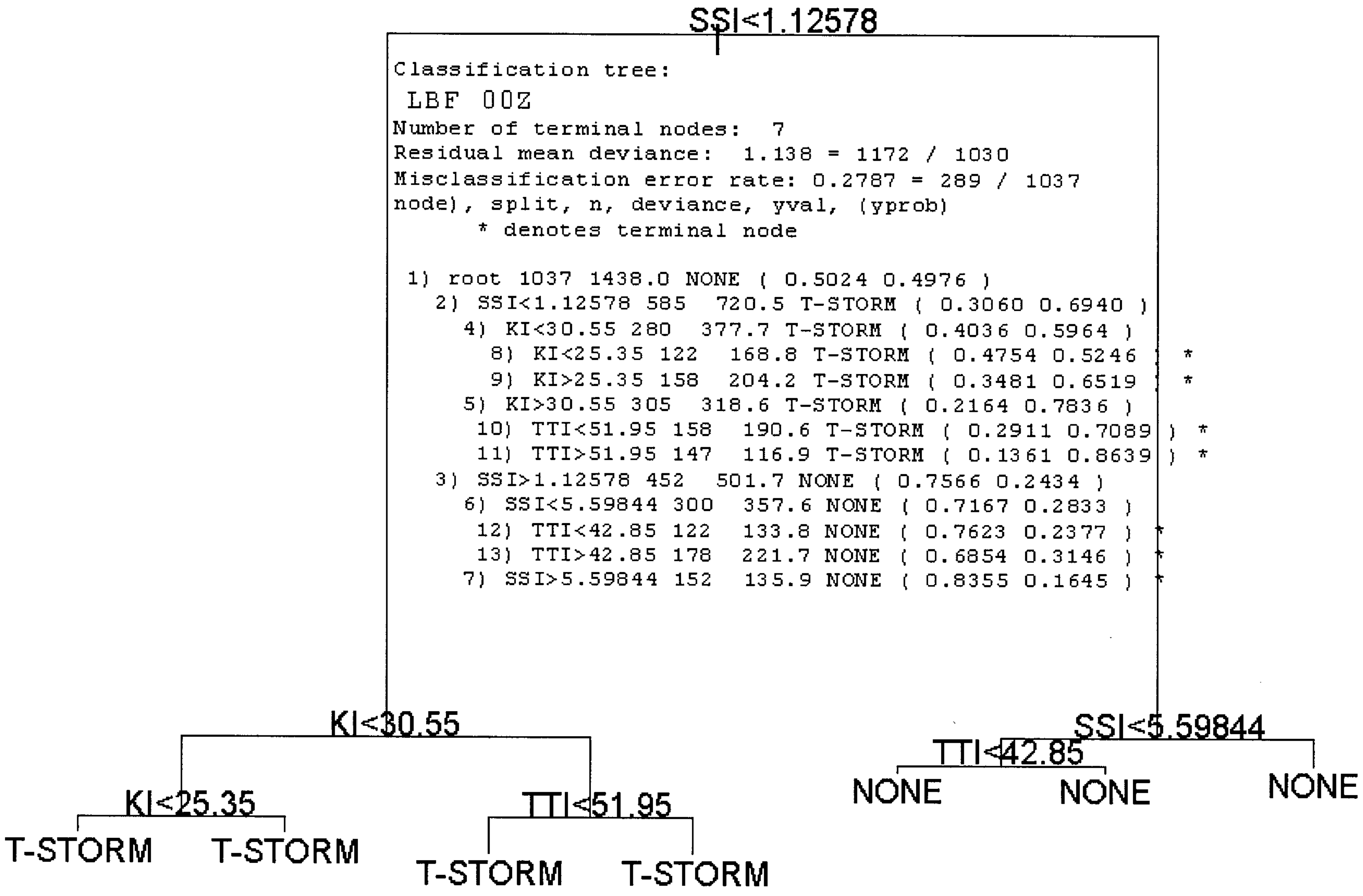
```
Regression tree:
DDC 00Z
Number of terminal nodes:  4
Residual mean deviance:  3706000 = 1901000000 / 513
Distribution of residuals:
  Min. 1st Qu. Median        Mean 3rd Qu.  Max.
 -2007  -872.2 -405.6 -9.763e-014   164.5 10630
node), split, n, deviance, yval
      * denotes terminal node

 1) root 517 2085000000 1045.0
   2) LI<-4.32695 123 1064000000 2008.0 *
   3) LI>-4.32695 394  871400000  744.6
     6) SSI<-0.523438 246  747500000  947.9
      12) TTI<50.05 104  435400000 1153.0 *
      13) TTI>50.05 142  304500000  797.5 *
     7) SSI>-0.523438 148   96820000  406.6 *
```



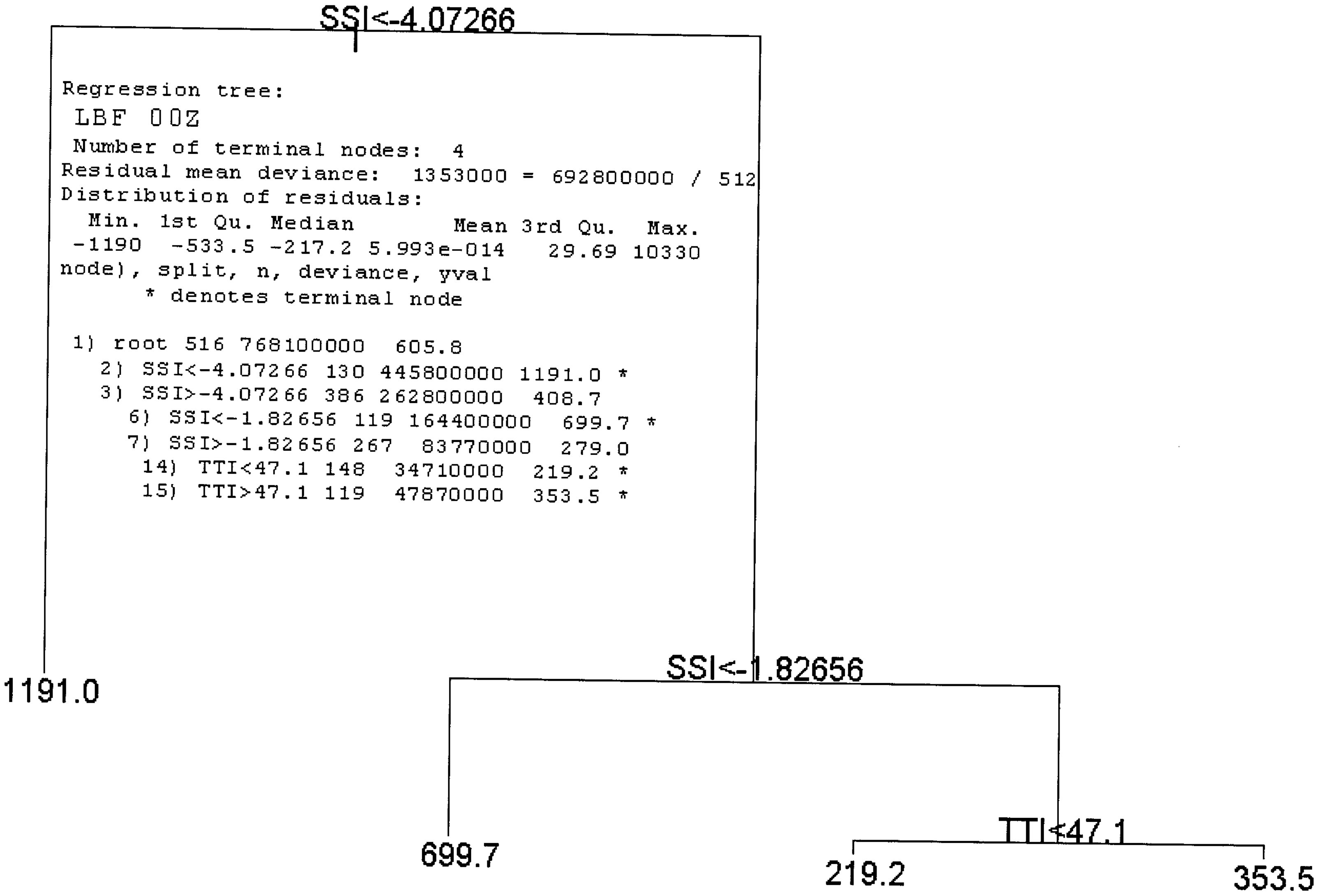
```
Classification tree:
 OUN 00Z
Number of terminal nodes:  8
Residual mean deviance:  1.219 = 1220 / 1001
Misclassification error rate: 0.33 = 333 / 1009
 1) root 1009 1381.00 NO t-storm ( 0.5669 0.4331 )
   2) KI<25.15 404  454.40 NO t-storm ( 0.7500 0.2500 )
     4) TTI<46.7 293  294.30 NO t-storm ( 0.7986 0.2014 )
       8) KI<10.8 107   83.03 NO t-storm ( 0.8692 0.1308 ) *
       9) KI>10.8 186  205.80 NO t-storm ( 0.7581 0.2419 ) *
     5) TTI>46.7 111  147.20 NO t-storm ( 0.6216 0.3784 ) *
   3) KI>25.15 605  831.30 t-storm ( 0.4446 0.5554 )
     6) LI<-1.06484 402  530.10 t-storm ( 0.3706 0.6294 )
      12) KI<35 245  336.20 t-storm ( 0.4408 0.5592 )
        24) SSI<-2.71875 103  136.70 t-storm ( 0.3786 0.6214 ) *
        25) SSI>-2.71875 142  196.70 t-storm ( 0.4859 0.5141 ) *
      13) KI>35 157  180.30 t-storm ( 0.2611 0.7389 ) *
     7) LI>-1.06484 203  274.60 NO t-storm ( 0.5911 0.4089 )
      14) SSI<0.814844 100  138.50 NO t-storm ( 0.5200 0.4800 ) *
      15) SSI>0.814844 103  132.00 NO t-storm ( 0.6602 0.3398 ) *
```



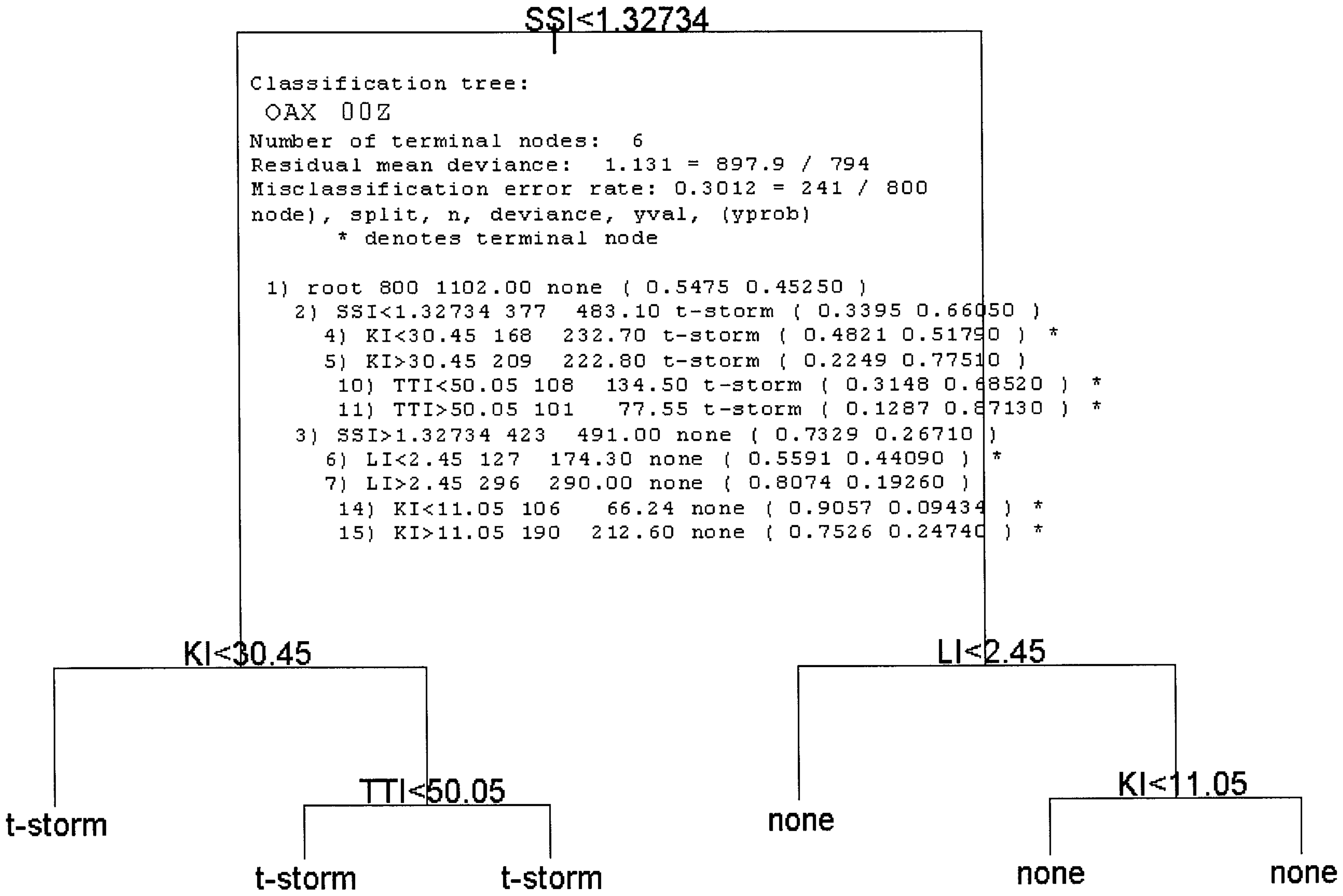
```
Regression tree:
OUN 00Z
Number of terminal nodes:  4
Residual mean deviance:  2787000 = 1207000000 / 433
Distribution of residuals:
  Min. 1st Qu. Median       Mean 3rd Qu.  Max.
 -1454  -630.8 -226.4 -1.93e-013  -39.43 19830
node), split, n, deviance, yval
      * denotes terminal node

 1) root 437 1299000000  678.3
   2) LI<-4.30898 103  886400000 1455.0 *
   3) LI>-4.30898 334  331000000  438.9
     6) TTI<45.75 129   40410000  227.4 *
     7) TTI>45.75 205  281200000  571.9
      14) SSI<-1.78203 103   90990000  491.8 *
      15) SSI>-1.78203 102  188800000  652.8 *
```



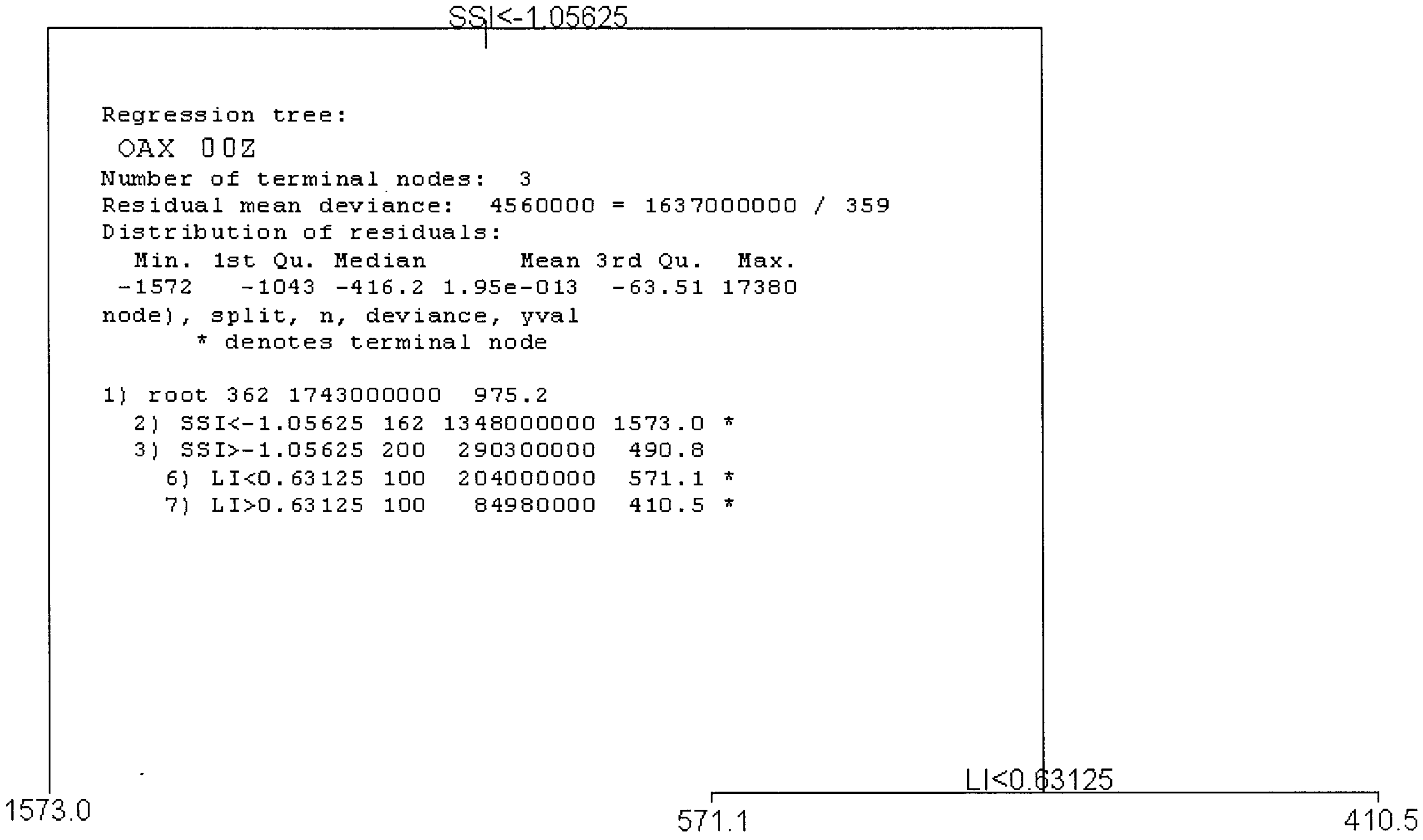
```
Classification tree:
 LBF 00Z
Number of terminal nodes:  7
Residual mean deviance:  1.138 = 1172 / 1030
Misclassification error rate: 0.2787 = 289 / 1037
node), split, n, deviance, yval, (yprob)
      * denotes terminal node

 1) root 1037 1438.0 NONE ( 0.5024 0.4976 )
   2) SSI<1.12578 585  720.5 T-STORM ( 0.3060 0.6940 )
     4) KI<30.55 280  377.7 T-STORM ( 0.4036 0.5964 )
       8) KI<25.35 122  168.8 T-STORM ( 0.4754 0.5246 ) *
       9) KI>25.35 158  204.2 T-STORM ( 0.3481 0.6519 ) *
     5) KI>30.55 305  318.6 T-STORM ( 0.2164 0.7836 )
      10) TTI<51.95 158  190.6 T-STORM ( 0.2911 0.7089 ) *
      11) TTI>51.95 147  116.9 T-STORM ( 0.1361 0.8639 ) *
   3) SSI>1.12578 452  501.7 NONE ( 0.7566 0.2434 )
     6) SSI<5.59844 300  357.6 NONE ( 0.7167 0.2833 )
      12) TTI<42.85 122  133.8 NONE ( 0.7623 0.2377 ) *
      13) TTI>42.85 178  221.7 NONE ( 0.6854 0.3146 ) *
     7) SSI>5.59844 152  135.9 NONE ( 0.8355 0.1645 ) *
```



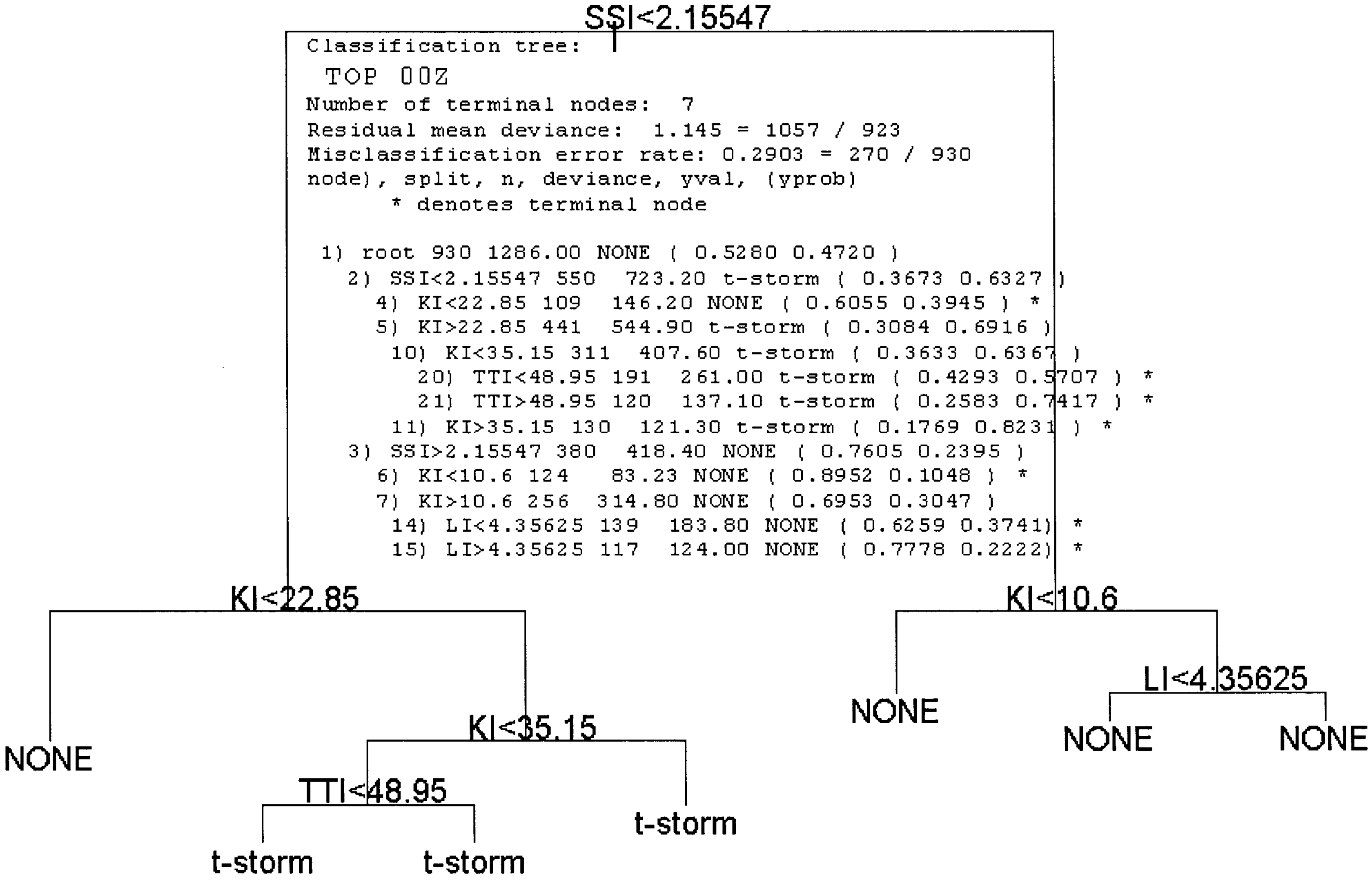
```
Regression tree:
 LBF 00Z
 Number of terminal nodes:  4
Residual mean deviance:  1353000 = 692800000 / 512
Distribution of residuals:
  Min. 1st Qu. Median        Mean 3rd Qu.  Max.
 -1190  -533.5 -217.2 5.993e-014   29.69 10330
node), split, n, deviance, yval
      * denotes terminal node

 1) root 516 768100000  605.8
   2) SSI<-4.07266 130 445800000 1191.0 *
   3) SSI>-4.07266 386 262800000  408.7
     6) SSI<-1.82656 119 164400000  699.7 *
     7) SSI>-1.82656 267  83770000  279.0
      14) TTI<47.1 148  34710000  219.2 *
      15) TTI>47.1 119  47870000  353.5 *
```



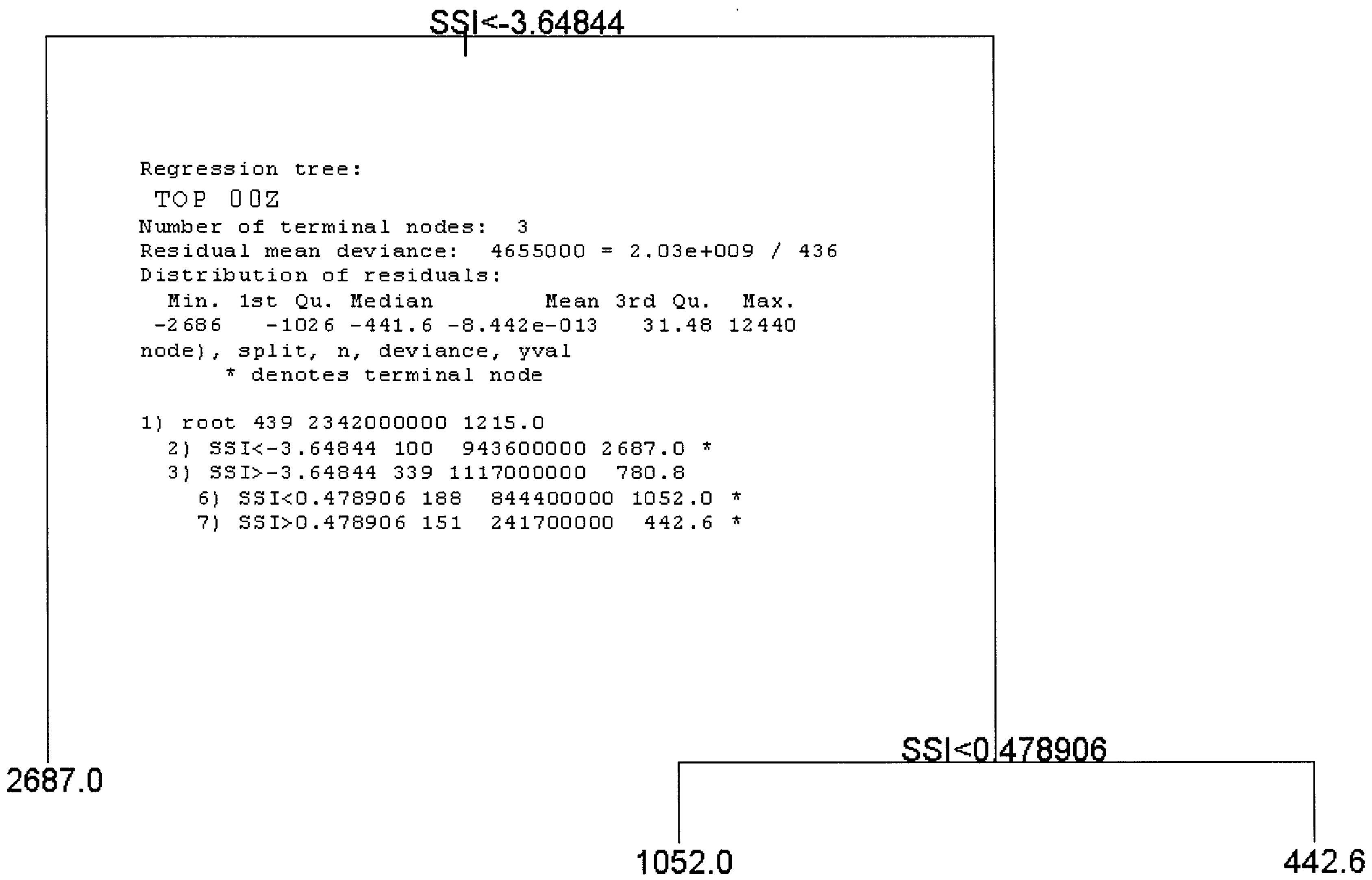
```
Classification tree:
 OAX 00Z
Number of terminal nodes:  6
Residual mean deviance:  1.131 = 897.9 / 794
Misclassification error rate: 0.3012 = 241 / 800
node), split, n, deviance, yval, (yprob)
      * denotes terminal node

 1) root 800 1102.00 none ( 0.5475 0.45250 )
   2) SSI<1.32734 377  483.10 t-storm ( 0.3395 0.66050 )
     4) KI<30.45 168  232.70 t-storm ( 0.4821 0.51790 ) *
     5) KI>30.45 209  222.80 t-storm ( 0.2249 0.77510 )
      10) TTI<50.05 108  134.50 t-storm ( 0.3148 0.68520 ) *
      11) TTI>50.05 101   77.55 t-storm ( 0.1287 0.87130 ) *
   3) SSI>1.32734 423  491.00 none ( 0.7329 0.26710 )
     6) LI<2.45 127  174.30 none ( 0.5591 0.44090 ) *
     7) LI>2.45 296  290.00 none ( 0.8074 0.19260 )
      14) KI<11.05 106   66.24 none ( 0.9057 0.09434 ) *
      15) KI>11.05 190  212.60 none ( 0.7526 0.24740 ) *
```



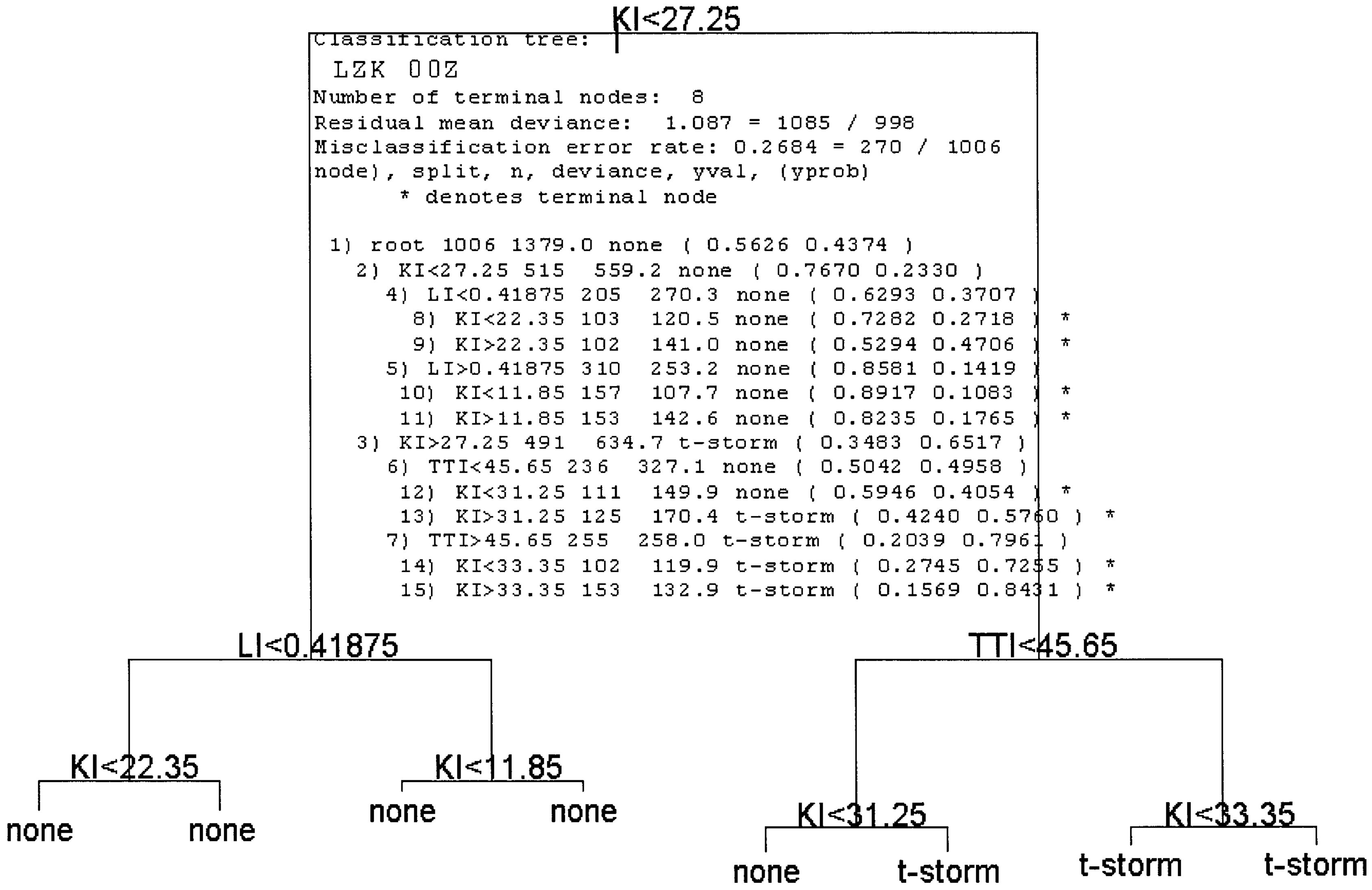
```
Regression tree:
 OAX 00Z
Number of terminal nodes:  3
Residual mean deviance:  4560000 = 1637000000 / 359
Distribution of residuals:
  Min. 1st Qu. Median       Mean 3rd Qu.  Max.
 -1572   -1043 -416.2 1.95e-013  -63.51 17380
node), split, n, deviance, yval
      * denotes terminal node

1) root 362 1743000000  975.2
  2) SSI<-1.05625 162 1348000000 1573.0 *
  3) SSI>-1.05625 200  290300000  490.8
    6) LI<0.63125 100  204000000  571.1 *
    7) LI>0.63125 100   84980000  410.5 *
```



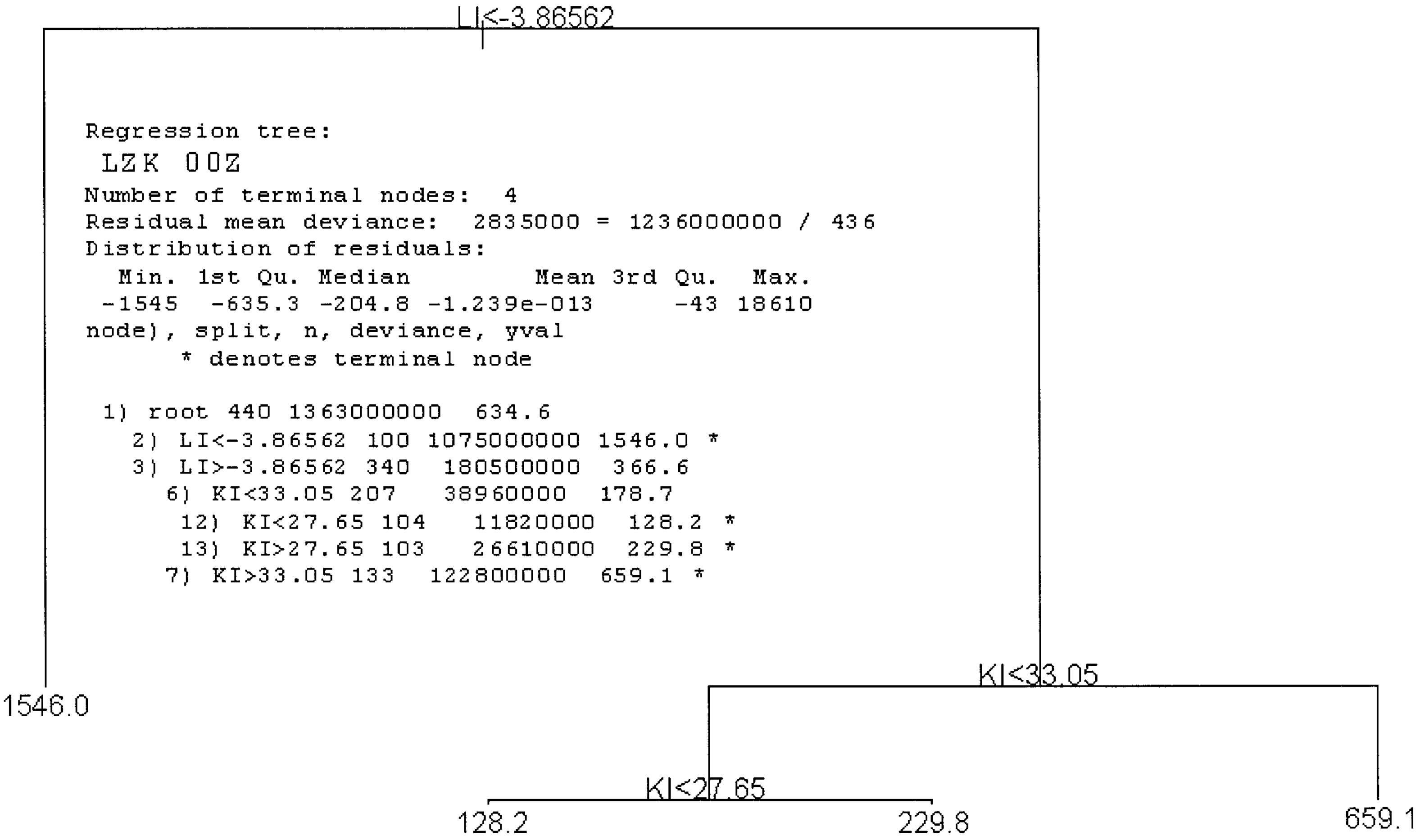
```
Classification tree:
 TOP 00Z
Number of terminal nodes:  7
Residual mean deviance:  1.145 = 1057 / 923
Misclassification error rate: 0.2903 = 270 / 930
node), split, n, deviance, yval, (yprob)
      * denotes terminal node

 1) root 930 1286.00 NONE ( 0.5280 0.4720 )
   2) SSI<2.15547 550  723.20 t-storm ( 0.3673 0.6327 )
     4) KI<22.85 109  146.20 NONE ( 0.6055 0.3945 ) *
     5) KI>22.85 441  544.90 t-storm ( 0.3084 0.6916 )
      10) KI<35.15 311  407.60 t-storm ( 0.3633 0.6367 )
        20) TTI<48.95 191  261.00 t-storm ( 0.4293 0.5707 ) *
        21) TTI>48.95 120  137.10 t-storm ( 0.2583 0.7417 ) *
      11) KI>35.15 130  121.30 t-storm ( 0.1769 0.8231 ) *
   3) SSI>2.15547 380  418.40 NONE ( 0.7605 0.2395 )
     6) KI<10.6 124   83.23 NONE ( 0.8952 0.1048 ) *
     7) KI>10.6 256  314.80 NONE ( 0.6953 0.3047 )
      14) LI<4.35625 139  183.80 NONE ( 0.6259 0.3741) *
      15) LI>4.35625 117  124.00 NONE ( 0.7778 0.2222) *
```



```
Regression tree:
 TOP 00Z
Number of terminal nodes:  3
Residual mean deviance:  4655000 = 2.03e+009 / 436
Distribution of residuals:
  Min. 1st Qu. Median        Mean 3rd Qu.  Max.
 -2686   -1026 -441.6 -8.442e-013   31.48 12440
node), split, n, deviance, yval
      * denotes terminal node

1) root 439 2342000000 1215.0
  2) SSI<-3.64844 100  943600000 2687.0 *
  3) SSI>-3.64844 339 1117000000  780.8
    6) SSI<0.478906 188  844400000 1052.0 *
    7) SSI>0.478906 151  241700000  442.6 *
```

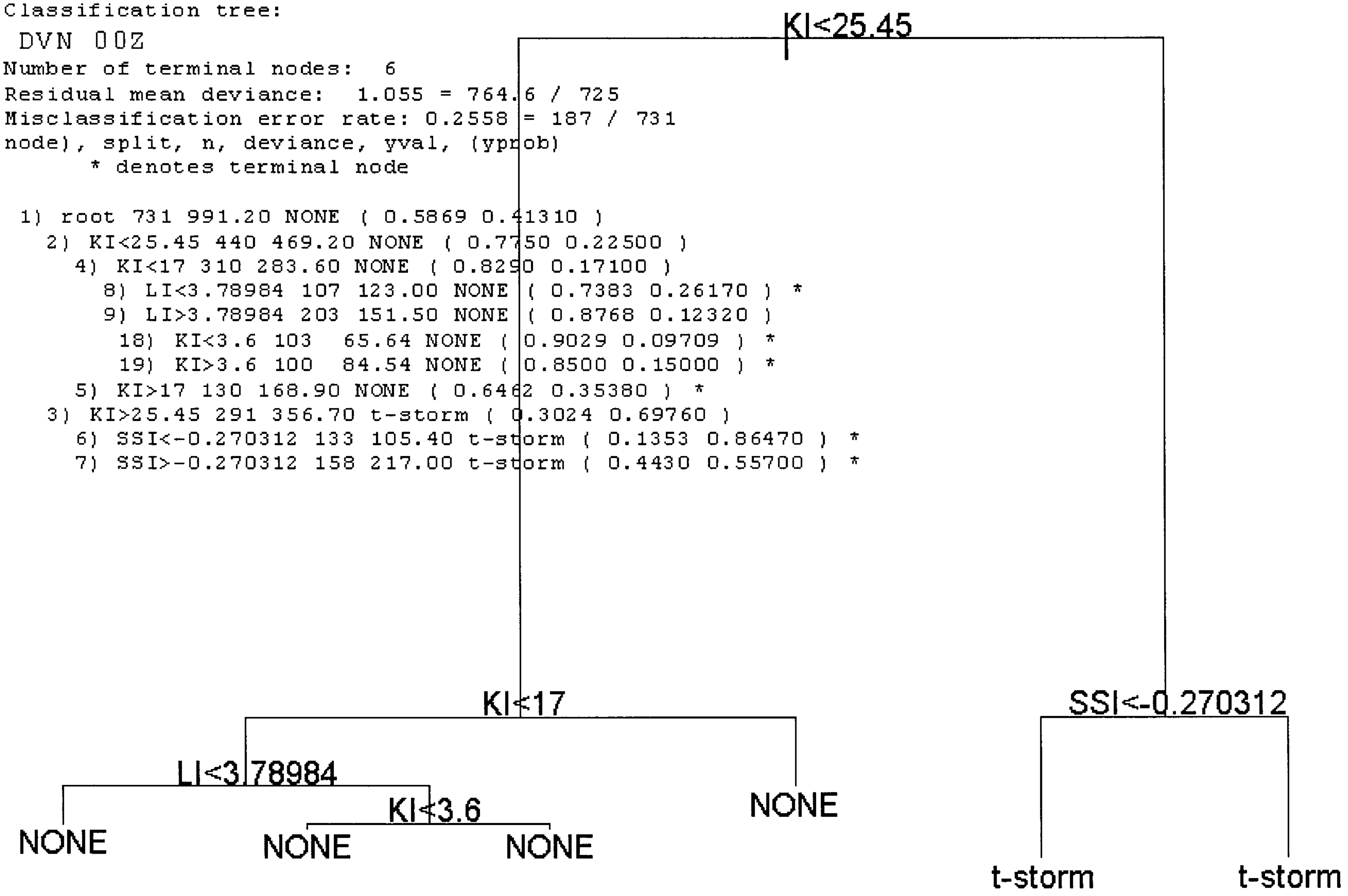


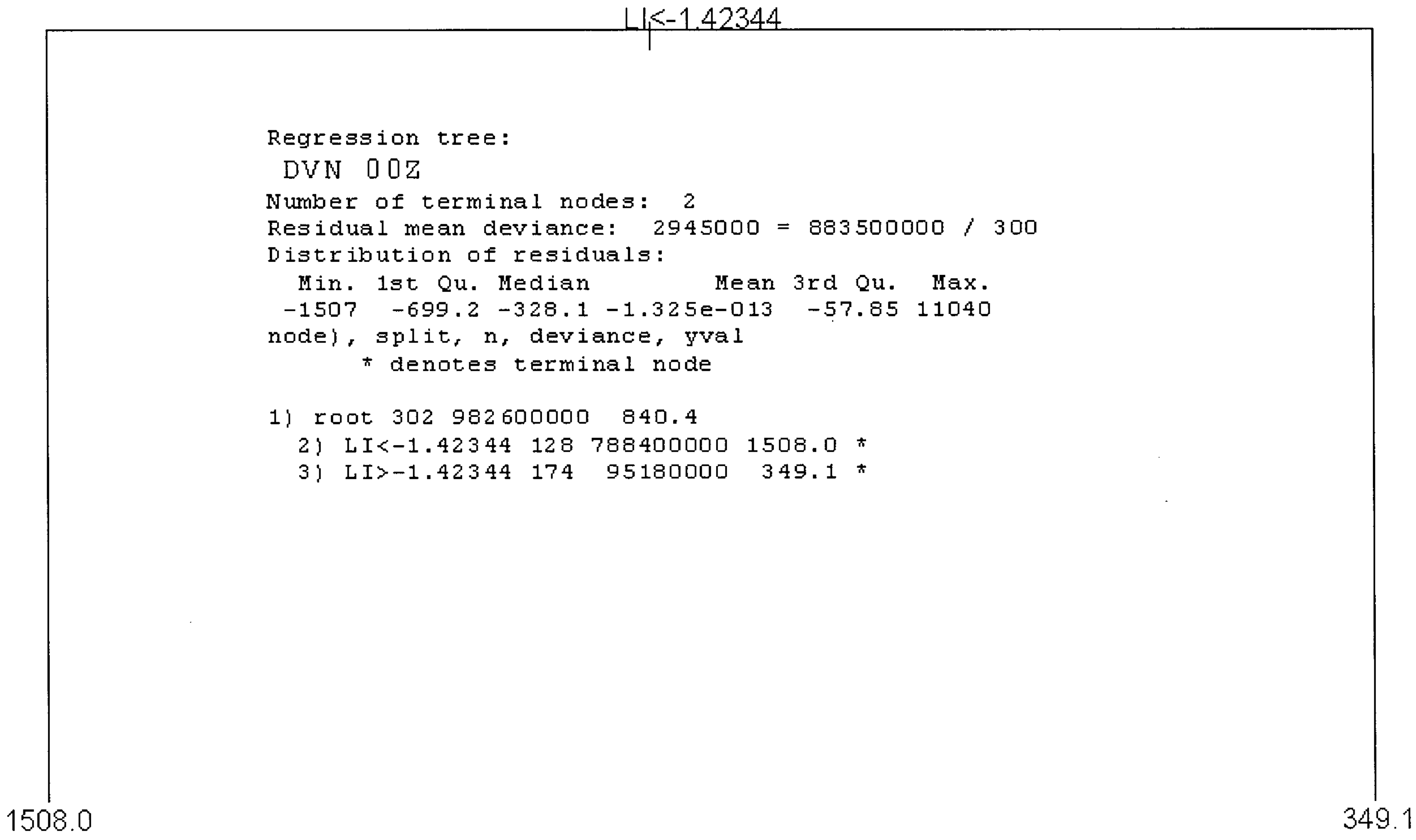
```
Classification tree:
 LZK 00Z
Number of terminal nodes:  8
Residual mean deviance:  1.087 = 1085 / 998
Misclassification error rate: 0.2684 = 270 / 1006
node), split, n, deviance, yval, (yprob)
      * denotes terminal node

 1) root 1006 1379.0 none ( 0.5626 0.4374 )
   2) KI<27.25 515  559.2 none ( 0.7670 0.2330 )
     4) LI<0.41875 205  270.3 none ( 0.6293 0.3707 )
       8) KI<22.35 103  120.5 none ( 0.7282 0.2718 ) *
       9) KI>22.35 102  141.0 none ( 0.5294 0.4706 ) *
     5) LI>0.41875 310  253.2 none ( 0.8581 0.1419 )
      10) KI<11.85 157  107.7 none ( 0.8917 0.1083 ) *
      11) KI>11.85 153  142.6 none ( 0.8235 0.1765 ) *
   3) KI>27.25 491  634.7 t-storm ( 0.3483 0.6517 )
     6) TTI<45.65 236  327.1 none ( 0.5042 0.4958 )
      12) KI<31.25 111  149.9 none ( 0.5946 0.4054 ) *
      13) KI>31.25 125  170.4 t-storm ( 0.4240 0.5760 ) *
     7) TTI>45.65 255  258.0 t-storm ( 0.2039 0.7961 )
      14) KI<33.35 102  119.9 t-storm ( 0.2745 0.7255 ) *
      15) KI>33.35 153  132.9 t-storm ( 0.1569 0.8431 ) *
```



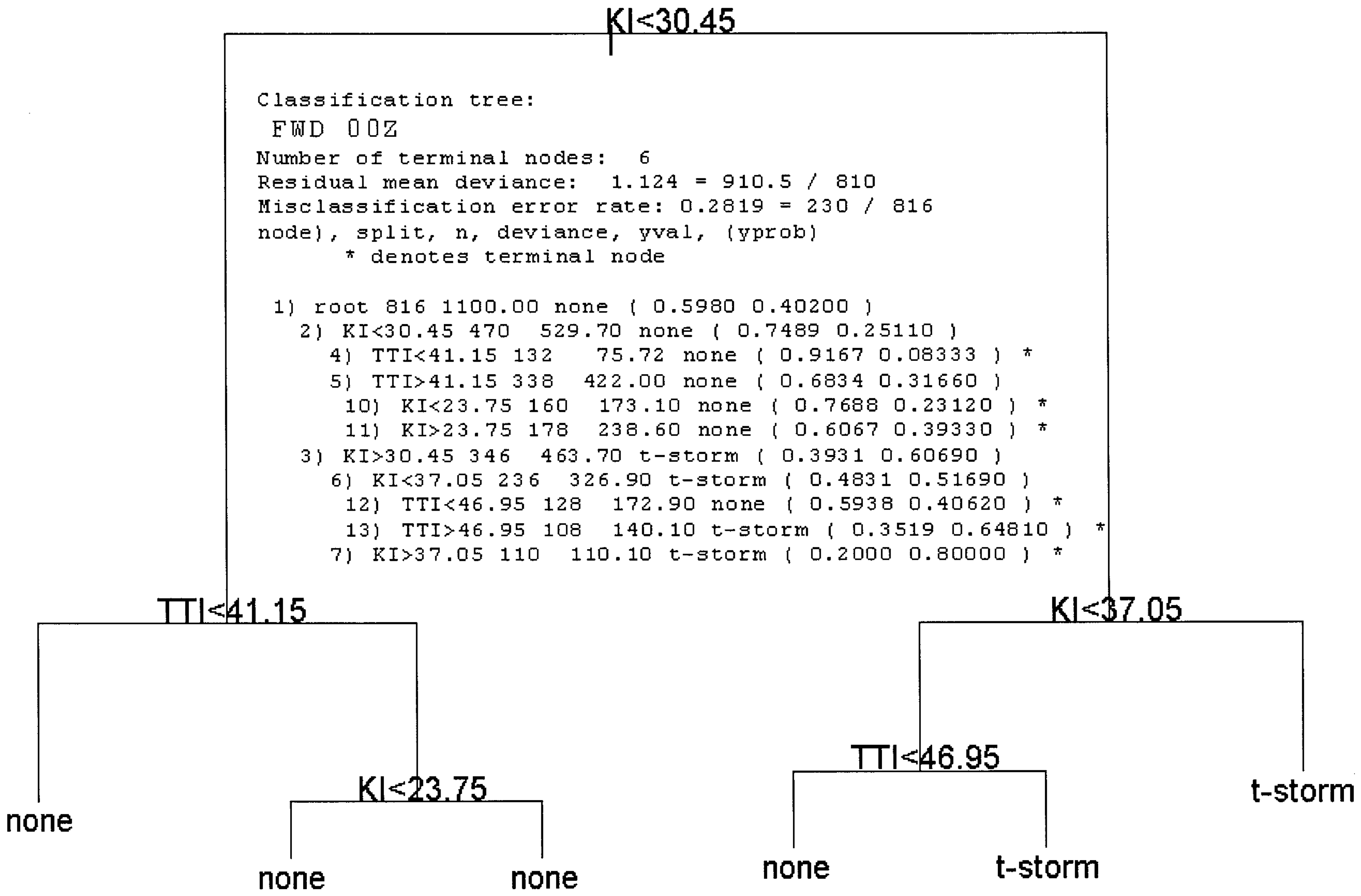
```
Regression tree:
 LZK 00Z
Number of terminal nodes:  4
Residual mean deviance:  2835000 = 1236000000 / 436
Distribution of residuals:
  Min. 1st Qu. Median        Mean 3rd Qu.  Max.
 -1545  -635.3 -204.8 -1.239e-013     -43 18610
node), split, n, deviance, yval
      * denotes terminal node

 1) root 440 1363000000  634.6
   2) LI<-3.86562 100 1075000000 1546.0 *
   3) LI>-3.86562 340  180500000  366.6
     6) KI<33.05 207   38960000  178.7
      12) KI<27.65 104   11820000  128.2 *
      13) KI>27.65 103   26610000  229.8 *
     7) KI>33.05 133  122800000  659.1 *
```

```
Classification tree:
 DVN 00Z
Number of terminal nodes:  6
Residual mean deviance:  1.055 = 764.6 / 725
Misclassification error rate: 0.2558 = 187 / 731
node), split, n, deviance, yval, (yprob)
      * denotes terminal node

 1) root 731 991.20 NONE ( 0.5869 0.41310 )
   2) KI<25.45 440 469.20 NONE ( 0.7750 0.22500 )
     4) KI<17 310 283.60 NONE ( 0.8290 0.17100 )
       8) LI<3.78984 107 123.00 NONE ( 0.7383 0.26170 ) *
       9) LI>3.78984 203 151.50 NONE ( 0.8768 0.12320 )
        18) KI<3.6 103  65.64 NONE ( 0.9029 0.09709 ) *
        19) KI>3.6 100  84.54 NONE ( 0.8500 0.15000 ) *
     5) KI>17 130 168.90 NONE ( 0.6462 0.35380 ) *
   3) KI>25.45 291 356.70 t-storm ( 0.3024 0.69760 )
     6) SSI<-0.270312 133 105.40 t-storm ( 0.1353 0.86470 ) *
     7) SSI>-0.270312 158 217.00 t-storm ( 0.4430 0.55700 ) *
```

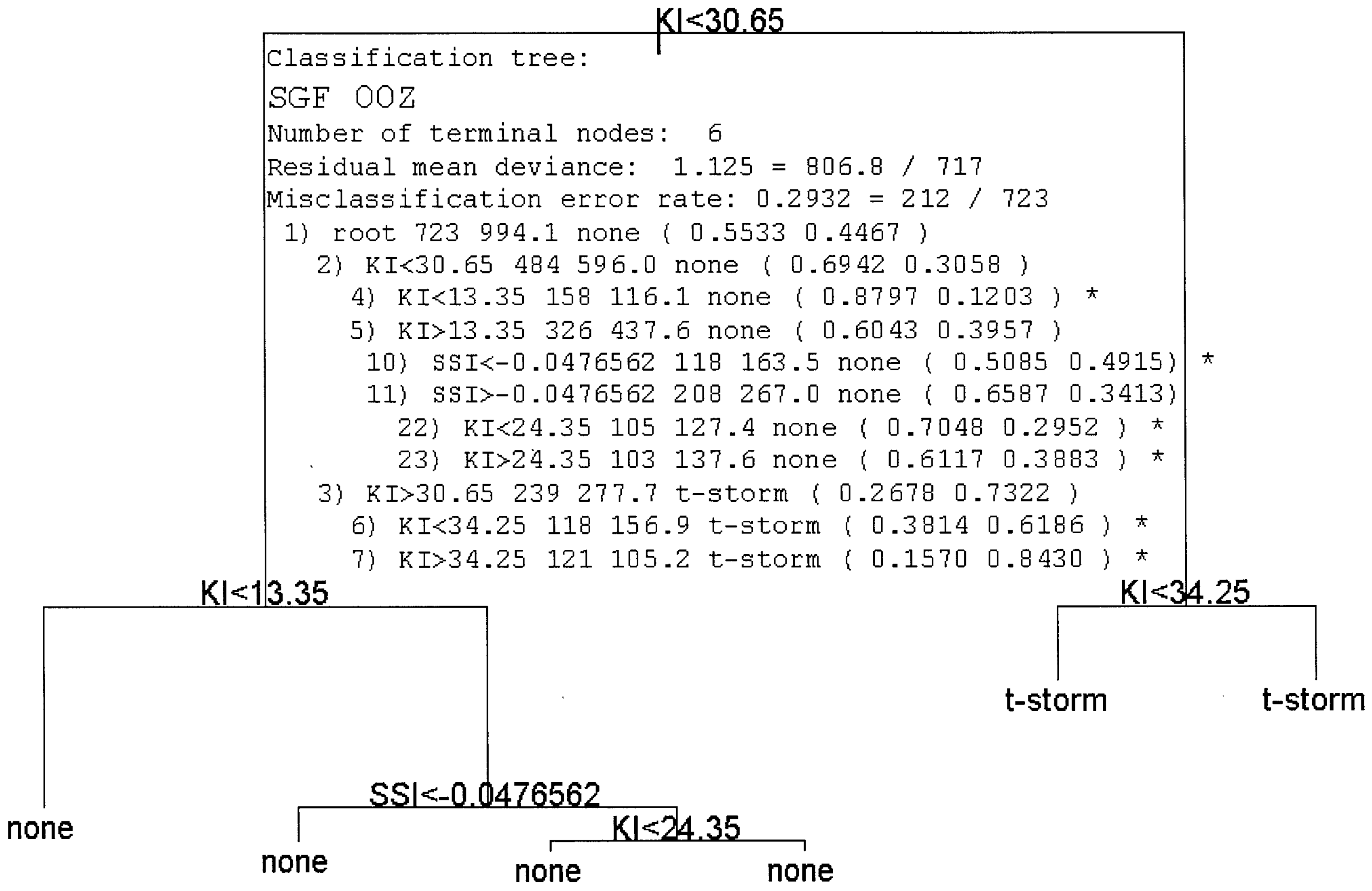




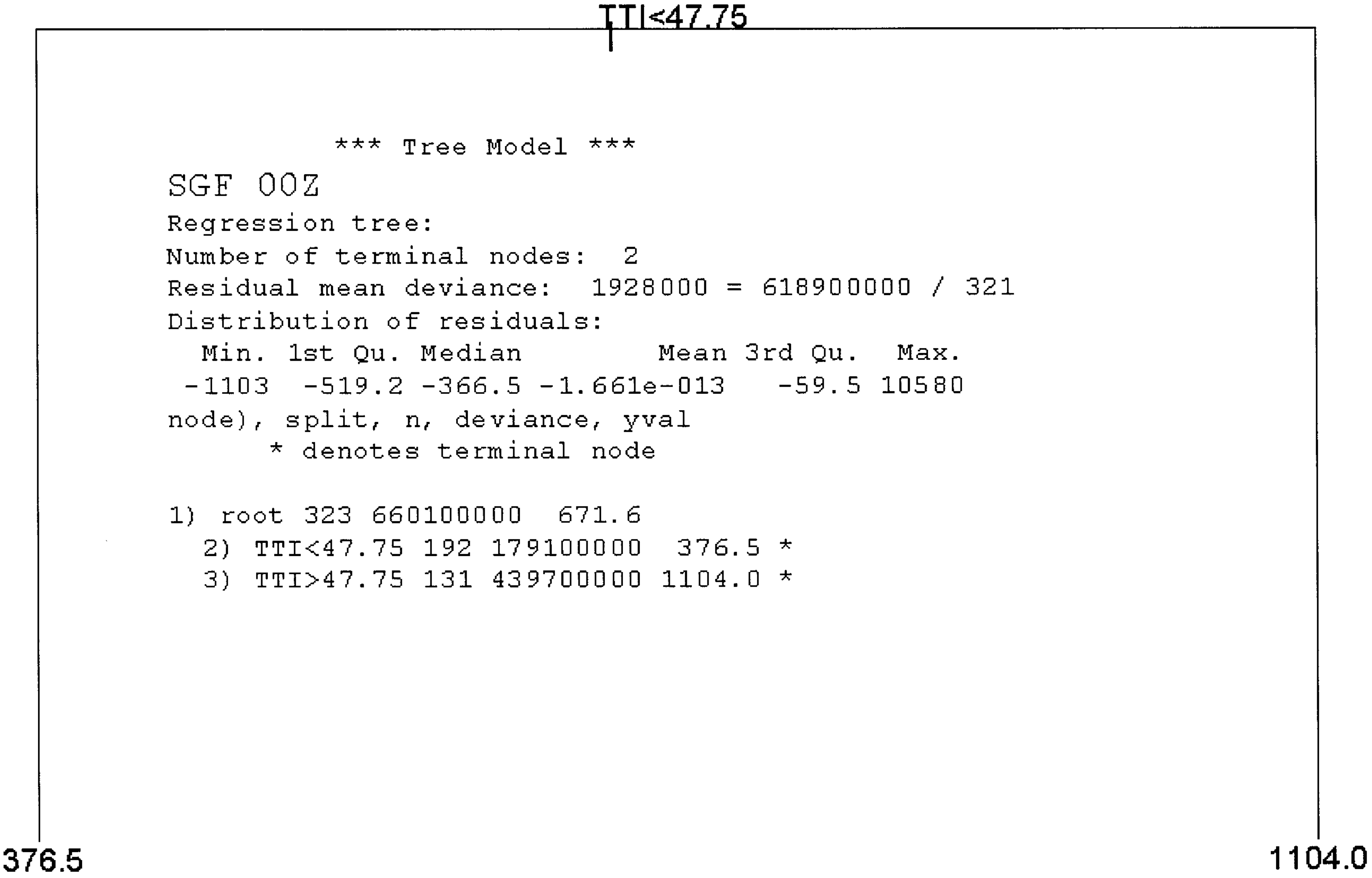
```
Regression tree:
 DVN 00Z
Number of terminal nodes:  2
Residual mean deviance:  2945000 = 883500000 / 300
Distribution of residuals:
  Min. 1st Qu. Median        Mean 3rd Qu.  Max.
 -1507  -699.2 -328.1 -1.325e-013  -57.85 11040
node), split, n, deviance, yval
      * denotes terminal node

1) root 302 982600000  840.4
  2) LI<-1.42344 128 788400000 1508.0 *
  3) LI>-1.42344 174  95180000  349.1 *
```



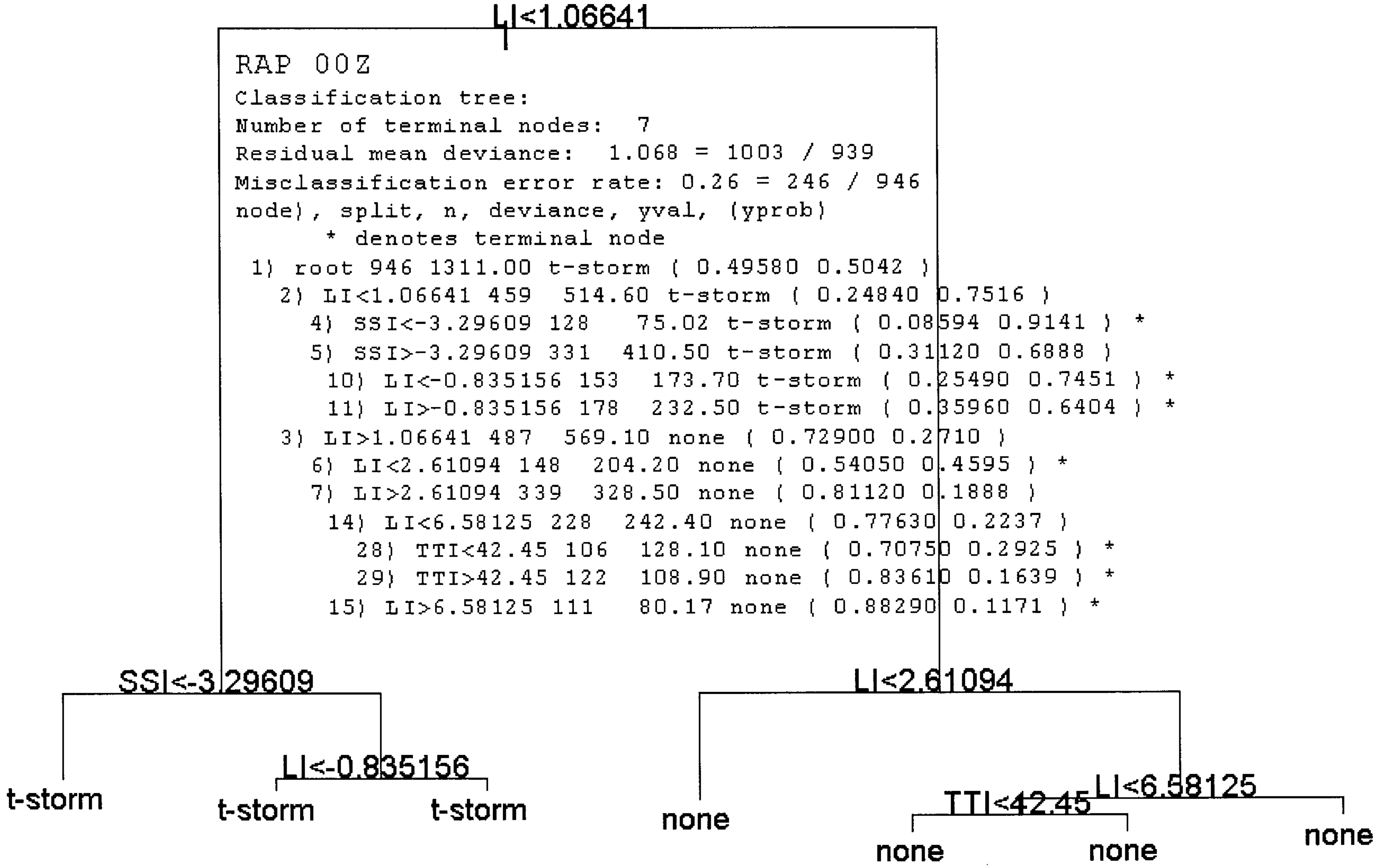
```
Classification tree:
 FWD 00Z
Number of terminal nodes:  6
Residual mean deviance:  1.124 = 910.5 / 810
Misclassification error rate: 0.2819 = 230 / 816
node), split, n, deviance, yval, (yprob)
      * denotes terminal node

 1) root 816 1100.00 none ( 0.5980 0.40200 )
   2) KI<30.45 470  529.70 none ( 0.7489 0.25110 )
     4) TTI<41.15 132   75.72 none ( 0.9167 0.08333 ) *
     5) TTI>41.15 338  422.00 none ( 0.6834 0.31660 )
      10) KI<23.75 160  173.10 none ( 0.7688 0.23120 ) *
      11) KI>23.75 178  238.60 none ( 0.6067 0.39330 ) *
   3) KI>30.45 346  463.70 t-storm ( 0.3931 0.60690 )
     6) KI<37.05 236  326.90 t-storm ( 0.4831 0.51690 )
      12) TTI<46.95 128  172.90 none ( 0.5938 0.40620 ) *
      13) TTI>46.95 108  140.10 t-storm ( 0.3519 0.64810 ) *
     7) KI>37.05 110  110.10 t-storm ( 0.2000 0.80000 ) *
```



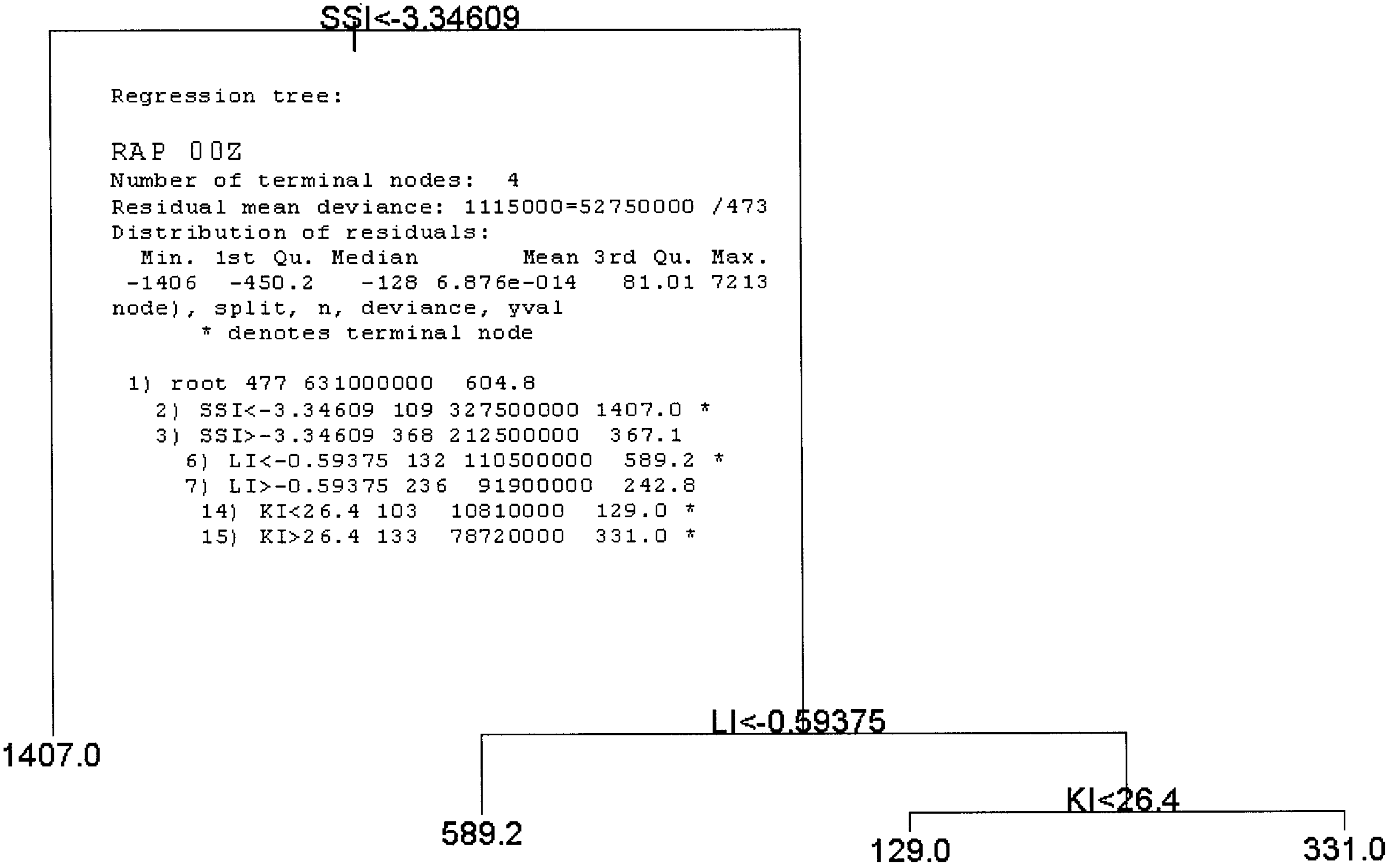
```
Classification tree:
SGF 00Z
Number of terminal nodes:  6
Residual mean deviance:  1.125 = 806.8 / 717
Misclassification error rate: 0.2932 = 212 / 723
 1) root 723 994.1 none ( 0.5533 0.4467 )
   2) KI<30.65 484 596.0 none ( 0.6942 0.3058 )
     4) KI<13.35 158 116.1 none ( 0.8797 0.1203 ) *
     5) KI>13.35 326 437.6 none ( 0.6043 0.3957 )
      10) SSI<-0.0476562 118 163.5 none ( 0.5085 0.4915) *
      11) SSI>-0.0476562 208 267.0 none ( 0.6587 0.3413)
        22) KI<24.35 105 127.4 none ( 0.7048 0.2952 ) *
        23) KI>24.35 103 137.6 none ( 0.6117 0.3883 ) *
   3) KI>30.65 239 277.7 t-storm ( 0.2678 0.7322 )
     6) KI<34.25 118 156.9 t-storm ( 0.3814 0.6186 ) *
     7) KI>34.25 121 105.2 t-storm ( 0.1570 0.8430 ) *
```



```
          *** Tree Model ***
SGF 00Z
Regression tree:
Number of terminal nodes:  2
Residual mean deviance:  1928000 = 618900000 / 321
Distribution of residuals:
  Min. 1st Qu. Median        Mean 3rd Qu.  Max.
 -1103  -519.2 -366.5 -1.661e-013   -59.5 10580
node), split, n, deviance, yval
      * denotes terminal node

1) root 323 660100000  671.6
  2) TTI<47.75 192 179100000  376.5 *
  3) TTI>47.75 131 439700000 1104.0 *
```



```
RAP 00Z
Classification tree:
Number of terminal nodes:  7
Residual mean deviance:  1.068 = 1003 / 939
Misclassification error rate: 0.26 = 246 / 946
node), split, n, deviance, yval, (yprob)
      * denotes terminal node
 1) root 946 1311.00 t-storm ( 0.49580 0.5042 )
   2) LI<1.06641 459  514.60 t-storm ( 0.24840 0.7516 )
     4) SSI<-3.29609 128   75.02 t-storm ( 0.08594 0.9141 ) *
     5) SSI>-3.29609 331  410.50 t-storm ( 0.31120 0.6888 )
      10) LI<-0.835156 153  173.70 t-storm ( 0.25490 0.7451 ) *
      11) LI>-0.835156 178  232.50 t-storm ( 0.35960 0.6404 ) *
   3) LI>1.06641 487  569.10 none ( 0.72900 0.2710 )
     6) LI<2.61094 148  204.20 none ( 0.54050 0.4595 ) *
     7) LI>2.61094 339  328.50 none ( 0.81120 0.1888 )
      14) LI<6.58125 228  242.40 none ( 0.77630 0.2237 )
        28) TTI<42.45 106  128.10 none ( 0.70750 0.2925 ) *
        29) TTI>42.45 122  108.90 none ( 0.83610 0.1639 ) *
      15) LI>6.58125 111   80.17 none ( 0.88290 0.1171 ) *
```



```
Regression tree:

RAP 00Z
Number of terminal nodes:  4
Residual mean deviance: 1115000=52750000 /473
Distribution of residuals:
  Min. 1st Qu. Median       Mean 3rd Qu. Max.
 -1406  -450.2   -128 6.876e-014   81.01 7213
node), split, n, deviance, yval
      * denotes terminal node

 1) root 477 631000000  604.8
   2) SSI<-3.34609 109 327500000 1407.0 *
   3) SSI>-3.34609 368 212500000  367.1
     6) LI<-0.59375 132 110500000  589.2 *
     7) LI>-0.59375 236  91900000  242.8
      14) KI<26.4 103  10810000  129.0 *
      15) KI>26.4 133  78720000  331.0 *
```

Bibliography

Air Force Weather Agency (AFWA), "Meteorological Techniques," Technical Note TN-98/002, 1998.

Air Weather Service (AWS) "The Use of the Skew-T, Log P Diagram in Analysis and Forecasting," Technical Reference TR 79-006 1990.

Andra, David. Science Operations Officer (SOO), National Weather Service, Norman, OK. Personal communication. 14 November 2000.

Bishop, Christopher M. Neural Networks for Pattern Recognition. *Oxford University Press*, 1995.

Blanchard, D. O. "Assessing the vertical distribution of Convective Available Potential Energy," *Wea. Forecasting*, 13, 870-877, 1998.

Breiman L., Friedman J.H., Olshen R.A., and Stone, C.J. (1984). Classification and Regression Trees. Wadsworth International Group, Belmont CA. Chambers, J.M., and Hastie, T.J., *Statistical Models in S*, pg. 414, 1991.

Brodley, C. E., Lane, T. and Stough, T., "Knowledge discovery and data mining," *American Scientist*, 87(1), 1999.

Byungyong, K. and D. Landgrebe. "Hierarchical decision tree classifiers in high-dimensional and large class data," *IEEE Transactions on Geoscience and Remote Sensing*, 29(4), 518-528, July 1991.

Coleman., "Thunderstorms Above Frontal Surfaces in Environments Without Positive CAPE Part I: A Climatology," *Mon. Wea. Rev.*, 118, 1103-1121, 1990.

_____________, "Thunderstorms Above Frontal Surfaces in Environments Without Positive CAPE Part II: Organization and Instability Mechanisms," *Mon. Wea. Rev.*, 118, 1123-1144, 1990.

Cummins K., M. J. Murphy, E.A. Bardo, W. L. Hiscox, R. B. Pyle, and A. E. Pifer. "A Combined TOA/MDF Technology Upgrade of the U.S. National Lightning Detection Network," *J. Geophys. Res.*, **103**, 9035-9044, 1998.

Dempsey, C., K. Howard, R. Maddox, D. Phillips, "Developing Advanced Weather Technologies for the Power Industry," *Bull. Amer. Met. Soc.*, 79, 1019-1036, 1998.

Devore, Jay L., *Probability and Statistics for Engineering and the Sciences*, USA: Wadsworth Publishing Co, 2000.

Dye, James E., Cloud Physics and Cloud Electrification What are the Connections? Preprints, *Conf. On Atmos. Electricity*, Amer. Meteor. Soc., Kananaskis Park Alberta, Canada, 687-691, 1990.

Fickett, J. and Tung, C.S., "Assessment of protein coding measures," *Nucleic Acids Res.*, 20, 6441-6450, 1992.

Friedman, J., "Data Mining and Statistics: What's the Connection?," Dept. of Statistics and Stanford Linear Accelerator Center, Stanford Univ., 1997.

Galway, J.G., The Lifted Index as a Predictor of Latent Instability. Bull. Amer. Meteor. Soc., 37, 528-529, 1956.

George, J.J., *Weather Forecasting for Aeronautics.* Academic Press, 673pp., 1960.

Huntrieser, H., H.H. Schiesser, W. Schmid, and A. Waldvogel, "Comparison of Traditional and Newly developed Thunderstorm Indices for Switzerland," *Wea. Forecasting*, 12, 108-123, 1996.

Koceilski, A. "A Preliminary Computer Analysis of Parameters Associated with Tornadic Activity," U.S. Weather Bureau Library, Kansas City, Mo.

Lyons W.A., Uliasz M., Nelson T. E., "Large Peak Current Cloud-to-Ground Lightning Flashes during the Summer Months in the Contiguous United States," *Monthly Weather Review, 126*, 2217-2233, 1998.

Marmelstein, Robert E. "*Evolving Compact Decision Rule Sets,*" PhD dissertation, Air Force Institute of Technology, Wright-Patterson AFB, OH, June 1999.

Miller, R.C., *Notes on analysis and severe storms forecasting procedures of the Air Force Global Weather Central*. Tech. Rep. 200 (rev.), Air Weather Service, 1972.

Mingers, J., "An empirical comparison of selection measures for Decision Tree Induction", Machine Learning, 3, 1989.

Mueller, C. K., J. W. Wilson, and N. A. Crook, "The utility of sounding and mesonet data to nowcast thunderstorm initiation," *Wea. Forecasting*, 8, 132-146, 1993.

Murthy, S., S. Kasif., S. Salzberg, and R. Beigel., "A system for induction of oblique decision trees," *Journal of Artificial Intelligence Research*, 2, 1-32, 1994.

NOAA, *Strategic Plan for Upper Air Observations*. 18pp., 1992.

Orville, R.E., and G.R. Huffines. "Lightning ground flash density over the contiguous United States," *Mon. Wea. Rev.*, 127, 2693-2703, 1999.

Rassmussen, E. and D. Blanchard, "A baseline climatology of sounding-derived supercell and tornado forecast parameters," Cooperative Institute for Mesoscale Meteorological Studies, National Severe Storms Lab and Univ. of Oklahoma, 1998.

Reap, R. M., and D. S. Foster. "Automated 12-36 hour probability forecasts of thunderstorms and severe local storms," *J. Appl. Meteor.*, 18. 1304-1315, 1979.

Rice, J. C. "Logistic regression: An introduction". B. Thompson, ed., *Advances in social science methodology*, Greenwich, CT: JAI Press, 3, 191-245, 1994.

Rymon, R. and N.M. Short, Jr. "Automatic cataloging and characterization of earch science data using set enumeration trees," *Telematics and Informatics*, 11(4), 309-318, Fall 1994.

Salzberg, S., R. Chandar, H. Ford, S. Murthy, and R. White. "Decision trees for automated identification of cosmic-ray hits in Hubble Space Telescope images," *Puplications of the Astronomical Society of the Pacific*, 107, 1-10, March 1995.

Salzberg, S. "Locating Protein Coding Regions in Human DNA using a Decision Tree," *Dissertation*, Dept. of Computer Science, Johns Hopkins Univ., 1995.

Showalter, A.K. "Stability Index for Forecasting Thunderstorms," *Bull. Amer. Met. Soc.*, 34(6), 250-252, June 1953.

Selvin, H. and A. Stuart. "Data Dredging procedures in survey analysis." *The American Statistician*, 20(3), 20-23, 1966.

South African Weather Bureau. "Background Information on instability indices and parameters related to Thunderstorm Development". Feb. 2000. http://www.weathersa.co.za/wfr/fcastaids/pcgrids/tsback.htm

Stuart, Neil A., Hugh D. Cobb, Wayne F. Albright, A. Todd Anderson, James Browder, Correlating Thunderstorm Lightning Patterns with WSR-88D Signatures and Resultant Benefits for Utility Companies: A Preliminary Investigation of the Hampton Roads "Hot Spot". Preprints, *19th Conference on Severe Local Storms*, Amer. Meteor. Soc., Minneapolis, Minnesota, In Press, 1998.

Wacker, R. S., and R. E. Orville, "Changes in measured lightning flash count and return stroke peak current after the 1994 U. S. National Lightning Detection Network upgrade," *J. Geophys. Res.*, 104, D2, 2151-2157, 1999.

Vita

Captain Kenneth C. Venzke was born on in Milwaukee, Wisconsin. He graduated from Saint Paul's College High School in Concordia, Missouri in May 1986 and soon after entered undergraduate studies at Colby College in Colby, Kansas where he graduated with a degree in Electronics in April 1989. He then entered his weather career in the Kansas Air National Guard where he first attended Basic Military Training School (BMT) in 1989. After graduating from the drum and bugle corps in BMT at Lackland AFB, TX he then attended Weather Observer school at Chanute AFB, IL and became a certified weather observer. In 1990 he transferred to the Meteorology department at the University of Kansas and graduated with a B.S. in Atmospheric Science in 1992. He then attended USAF Weather Forecaster School at Chanute AFB, IL for eight months. After a year with the National Weather Service in Kansas City he decided to be part of the action and serve his country full-time as an officer. He was commissioned in 1994 and was assigned to his first duty station at McConnell AFB, KS where his hard work and dedication to duty was recognized by receiving Company Grade Officer of the Year for the 22nd Operational Support Squadron. In March 1996, he was assigned to HQ Strategic Command, Offutt AFB, NE as a Battlestaff Weather Effects Officer aboard the "Looking Glass" Airborne Command Post (ABNCP). Currently he is attending the Graduate Meteorology program, Department of Engineering Physics, Air Force Institute of Technology. Upon graduation, he will be assigned to Keesler AFB, MS as a Weather Officer instructor.

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 074-0188*

1. REPORT DATE *(DD-MM-YYYY)* 01-03-2001	**2. REPORT TYPE** Master's Thesis	**3. DATES COVERED** *(From – To)* Jun 2000 – Sep 2001

4. TITLE AND SUBTITLE

Development of Predictors for Cloud-to-Ground Lightning Activity using Atmospheric Stability Indices

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Venzke, Kenneth C., Captain, USAF

5d. PROJECT NUMBER
If funded, enter ENR #

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S)

Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/EN)
2950 P Street, Building 640
WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT/GM/ENP/01M-8

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

AFCCC
Attn: Mr. Kenneth R. Walters
151 Patton Ave., Rm. 120, Federal Bldg.
Asheville, NC 28801 DSN: 673-9004

10. SPONSOR/MONITOR'S ACRONYM(S)

AFCCC/DOO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

A detailed examination was performed on several commonly applied atmospheric stability indices and lightning activity from 1993 to 2000 to determine the indices usefulness as predictive tools for determining cloud-to-ground lightning activity. Predetermined radii of 50 nautical miles around upper-air stations in the Midwest U.S. were used for the lightning summaries.

Also explored is an improvement upon the commonly accepted thresholds of the stability indices as general thunderstorm indicators. An improvement was found and new threshold ranges were developed for relating stability index values to lightning occurrence.

Traditional statistical regression methods failed to find a significant predictive relationship. By examining new techniques of data analysis, it was found that the detection and classification abilities of decision trees derived from the data-mining field best served the purposes of this study. Decision trees were examined on the large available database and significant results were found, resulting in the development of a lightning forecast tool for both the probability of lightning occurrence and its intensity. The predictive ability of the decision trees used in this study for lightning detection often exceeded 80-90% for most locations with a high degree of confidence.

The most significant features of the decision tree results were formulated into a forecast prediction tool with summary results for each location analyzed. These are specified both graphically and textually in a user-friendly format for forecasters to use as a "ready to use" predictive tool for forecasting lightning activity.
The results of this study using classification and regression trees were significant enough to implement immediately as a forecast tool for the operational weather forecast environment. Appendix A of this study is written as a "ready-to-use" forecast tool for weather forecasters. It is suggested that Air Force Weather units in the Midwest U.S. use this "innovative" forecast tool immediately for forecasting lightning activity.

15. SUBJECT TERMS

Stability Indices, Lightning Prediction, Thunderstorm Indicators, Data Mining, Lightning Forecast, Forecast Tool, Lightning Probability, Decision Tree, Classification Tree, Regression Tree, Cloud-to-Ground Lightning Activity, Stability Index Thresholds, Lifted Index, Showalter Index, SWEAT Index, KO Index, Total Totals Index, K-Index, CAPE

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Lt Col Ronald P. Lowther, ENP
a. REPOR T	**b. ABSTR ACT**	**c. THIS PAGE**			**19b. TELEPHONE NUMBER** *(Include area code)* (937) 255-3636, ext 4645
U	U	U	UU	197	

Form Approved
OMB No. 074-0188